

Diagnosing Knowledge Conflict in Multimodal Long-Chain Reasoning

Jing Tang^{1,*} and Kun Wang^{2,*} and Haolang Lu^{3,*} and Hongjin Chen³

KaiTao Chen³ and Zhongxiang Sun⁴ and Qiankun Li² and Lingjuan Lyu⁵

Guoshun Nan³ and Zhigang Zeng¹

¹Huazhong University of Science and Technology ²Nanyang Technological University

³Beijing University of Posts and Telecommunications ⁴Renmin University of China

⁵Sony AI, Zurich, Switzerland

jingtang@hust.edu.cn wang.kun@ntu.edu.sg luhaolang@bupt.edu.cn

*Equal contribution. Correspondence: nanguoshun@gmail.com, zgzeng@hust.edu.cn

Abstract

Multimodal large language models in long chain-of-thought reasoning often fail when different knowledge sources provide conflicting signals. We formalize these failures under a unified notion of knowledge conflict, distinguishing input-level objective conflict from process-level effective conflict. Through probing internal representations, we reveal that: (I) **Linear Separability**: different conflict types are explicitly encoded as linearly separable features rather than entangled; (II) **Depth Localization**: conflict signals concentrate in mid-to-late layers, indicating a distinct processing stage for conflict encoding; (III) **Hierarchical Consistency**: aggregating noisy token-level signals along trajectories robustly recovers input-level conflict types; and (IV) **Directional Asymmetry**: reinforcing the model’s implicit source preference under conflict is far easier than enforcing the opposite source. Our findings provide a representation-level characterization of multimodal reasoning under knowledge conflict and enable principled diagnosis and control of long-CoT failures. Code: github.com/Diagnosing-Knowledge-Conflict.

1 Introduction

Multimodal large language models (MLLMs) (Jin et al., 2025; Caffagni et al., 2024; Zhang et al., 2024a) have made substantial progress in visual understanding (Tong et al., 2024a; Ghatkesar et al., 2025; Ma et al., 2025), textual reasoning (Wang et al., 2024; Du et al., 2025a; Mirzadeh et al., 2025), and cross-modal alignment (Yu et al., 2024; Yan et al., 2025; Yu et al., 2025), enabling complex perception–reasoning–decision workflows. A defining capability is long-form reasoning: beyond producing answers, these models can generate extended chains-of-thought (CoT) (Wang et al., 2025b; Yue et al., 2025) that support challenging multi-step tasks. However, recent work documents failures under contradictory evidence or constraints: mod-

els may ignore explicit instructions (Wang et al., 2025a; Zhao et al., 2025), privilege the wrong evidential source (Guan et al., 2024; Liu et al., 2025b), or yield plausible yet goal-inconsistent conclusions (Fanous et al., 2025). These observations suggest that a **key bottleneck in multimodal reasoning is not missing information, but reliable decision-making under conflicting signals**.

Building on these observations, prior work (Zhang et al., 2024c; Lu et al., 2024) has characterized abnormal behavior under conflicting signals from several largely independent angles. In retrieval-augmented generation, a central question is whether models remain faithful to retrieved evidence or drift toward parametric priors (Wu et al., 2024). In vision settings with counterfactual or commonsense-violating inputs, MLLMs are often found to underweight visual evidence and default to “reasonable” answers that match world knowledge (Tong et al., 2024b; Liu et al., 2025c). In high-stakes domains, studies further report over-accommodation to user assertions, which can pull predictions away from the underlying evidence (Sharma et al., 2024). Although these lines of work differ in tasks, datasets, and evaluation criteria, their failure modes are similar: when information sources disagree, models do not reliably follow the appropriate basis for a decision, and instead exhibit unstable, hard-to-control trade-offs across sources.

In this paper, we take a unified view that these phenomena arise from *knowledge conflict* in multimodal reasoning. When generating tokens, MLLMs jointly rely on multiple knowledge sources, including visual evidence, textual instructions and contextual constraints, and parametric priors stored in the model weights (Han et al., 2026; Liu et al., 2024a; Karamcheti et al., 2024). When these sources provide inconsistent signals for the same goal, the model must resolve which source to follow. Importantly, the resulting failures are

not fabrications from missing knowledge, but incorrect source selection under conflict: **the model may have access to competing plausible cues yet follow the wrong basis**. Accordingly, our focus is not the act of answer generation itself, but whether conflict-induced failures can be *localized, measured, and functionally tested* at the level of internal representations.

Multimodal long-CoT reasoning (Ni et al., 2025) sharpens this problem: conflicts may arise at any step or modality, and once a step adopts the wrong basis, subsequent reasoning can proceed coherently yet end in a globally incorrect conclusion (Zhang et al., 2024b). Because such deviations are masked by fluent rationales (Turpin et al., 2023), the final output alone rarely reveals when a conflict emerged or how it propagated, so diagnosing long-CoT failures requires step-level tools.

In this work, ♣ We diagnose knowledge conflict dynamics on **7,500+** long-CoT trajectories from an *objective conflict* benchmark, where *effective conflicts* are activated in **78–90%** of samples; these rates are conditional on a constructed stress test and are not estimates of natural prevalence. ♣ Through layer-wise analysis of three models, we localize a depth-dependent *conflict-sensitive layer band*. Using streaming probes to detect token-level conflict states, we find they exhibit high linear separability (**93.2~98.8% AUC, 76.9~97.8% Recall@0.1**), revealing them as explicit, decodable features. ♣ We employ three *pluggable* methods for intervention. These methods can either steer model outputs toward selected directions, reducing conflict frequency by **up to 80%**, or suppress high-confidence errors by up to **55%**.

2 Related Work

Knowledge Conflict. Research on knowledge conflicts has identified three primary sources: conflicts between internal priors and visual information (Liu et al., 2025b; Du et al., 2025b) or textual inputs (Zhang et al., 2025a; Su et al., 2024), and conflicts between visual and textual modalities (Deng et al., 2025). Building on these findings, significant efforts have been made to mitigate such conflicts through advanced strategies (Xie et al., 2024; Guo et al., 2024), including knowledge editing (Tan et al., 2024; Zhang et al., 2025d; Cheng et al., 2024; Chen et al., 2025) and retrieval augmentation (Huo et al., 2025; Zhang et al., 2025b; Li et al., 2025). These approaches have demonstrated

potential in enhancing model faithfulness and reliability (Huang et al., 2025b; An et al., 2025; Shi et al., 2024; Zhang et al., 2024d; Lu et al., 2025). Text-only studies typically isolate a single source pair, most often context versus parametric memory (Xie et al., 2024; Jin et al., 2024), and the field survey (Xu et al., 2024) likewise organizes knowledge conflict along text-centric axes without modeling visual evidence. Although the above evidence suggests that conflicts are coupled and multi-source, existing solutions remain fragmented across modalities and fail to model conflicts holistically, thereby limiting their applicability in complex settings.

Probe Detection. Investigating internal states via probe detection is a developing field, yet the history of probing in LLMs (Kahana et al., 2025) provides clear precedents. Notably, the evolution of probe detection primarily centers on hallucination and faithfulness (Feng et al., 2025; Yi et al., 2025). Core techniques, such as linear probe generators (Kahana et al., 2025) and propositional probes (Feng et al., 2025), have inspired analogous approaches in watermark identification (Liu et al., 2025a), reward maximization (Li et al., 2024), and combinatorial optimization (Zhang et al., 2025e). However, these approaches predominantly focus on single-modal issues or specific downstream tasks, leaving the detection and localization of multimodal knowledge conflicts largely unexplored. Inspired by this, we introduce a specialized probe detection framework to identify the three sources of knowledge conflicts in MLLMs.

3 Conflict in Multimodal Reasoning

3.1 Knowledge Sources and Pairwise Conflicts

We consider a multimodal long-CoT reasoning task with input $x = (X_V, X_T)$ (visual X_V , textual X_T). Given a multimodal generative model M_θ , reasoning unfolds as a token sequence $\tau(x) = (y_1, \dots, y_T)$, with each token sampled as $y_t \sim M_\theta(\cdot \mid x, y_{<t})$ and the step- t state denoted $\mathbf{h}_t = f_\theta(x, y_{<t})$, where f_θ extracts the hidden representation via a forward pass up to a specified layer.

We abstract the knowledge available to the model into three sources, $\mathcal{K} = \{\mathcal{K}_{\text{vision}}, \mathcal{K}_{\text{text}}, \mathcal{K}_{\text{prior}}\}$: $\mathcal{K}_{\text{vision}}$ and $\mathcal{K}_{\text{text}}$ consist of facts supported by X_V and X_T respectively, while $\mathcal{K}_{\text{prior}}$ denotes parametric priors

Table 1: **Output-level conflict profile across models (objective conflict subsets).** We present statistics of generated trajectories under three types of conflict (model details in Appendix B). Metrics reported include sample count, average CoT length, average conflict spans per sample (*spans are contiguous conflict segments identified via an automated LLM annotation pipeline, may consist of one or multiple tokens.*), conflict token density (proportion of conflicting tokens), and sample conflict rate (% of samples exhibiting *effective conflict*).

Metric	Llama-3.2V-11B-cot				R1-Onevision-7B				Ocean-R1-7B-Instruct			
	$\mathcal{C}_{VP}^o = 1$	$\mathcal{C}_{VT}^o = 1$	$\mathcal{C}_{PT}^o = 1$	All	$\mathcal{C}_{VP}^o = 1$	$\mathcal{C}_{VT}^o = 1$	$\mathcal{C}_{PT}^o = 1$	All	$\mathcal{C}_{VP}^o = 1$	$\mathcal{C}_{VT}^o = 1$	$\mathcal{C}_{PT}^o = 1$	All
Samples	749	1012	803	2564	724	993	769	2486	640	1026	807	2473
Avg. CoT length (tokens)	326.79	1768.85	238.50	868.32	706.85	790.63	558.97	694.57	488.15	711.26	302.97	520.28
Avg. conflict spans per sample	2.69	6.20	4.04	4.50	3.66	6.73	7.02	5.93	8.68	9.00	5.43	7.75
Conflict token density (%)	4.92	1.65	11.25	5.61	3.20	2.16	7.68	4.17	8.70	3.23	11.77	7.43
Conflict Sample Ratio (%)	63.68	82.21	86.43	78.12	59.67	85.90	87.91	78.88	88.75	90.25	89.34	89.57

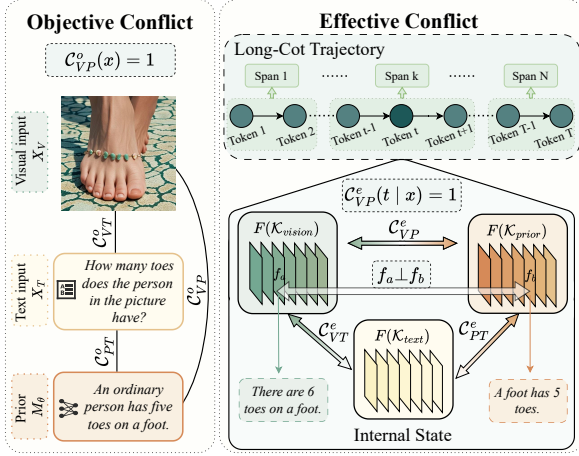


Figure 1: **Overview of Knowledge Sources and Conflict Types.** We categorize knowledge into Visual ($\mathcal{K}_{\text{vision}}$), Textual ($\mathcal{K}_{\text{text}}$), and Parametric Prior ($\mathcal{K}_{\text{prior}}$). Knowledge conflicts arise when factual statements from different sources act as incompatible signals. We define three primary conflict types: Vision-Text ($\mathcal{C}_{VT}^o$), Vision-Prior ($\mathcal{C}_{VP}^o$), and Prior-Text ($\mathcal{C}_{PT}^o$).

encoded in θ . For each $\mathcal{K}_* \in \mathcal{K}$, we represent its factual content as a set of atomic statements $F(\mathcal{K}_*)$, and write $\psi_a \perp \psi_b$ when two facts are semantically incompatible. We define a *pairwise knowledge conflict* between sources $\mathcal{K}_i$ and $\mathcal{K}_j$ ($i \neq j$) as the set of incompatible fact pairs:

$$\mathcal{C}_{i,j} = \{(\psi_i, \psi_j) \mid \psi_i \in F(\mathcal{K}_i), \psi_j \in F(\mathcal{K}_j), \psi_i \perp \psi_j\}. \quad (1)$$

We focus on three primary types induced by the three sources: Vision-Prior ($\mathcal{C}_{VP}^o$), Vision-Text ($\mathcal{C}_{VT}^o$), and Prior-Text ($\mathcal{C}_{PT}^o$).

3.2 Objective vs. Effective Conflict

As shown in Figure 1, we distinguish two notions: *objective conflict*, defined at the input level, and *effective conflict*, a process-level reasoning state.

Objective Conflict describes factual inconsistency induced by the input and the model’s parametric priors. For a conflict type $\mathcal{C}_{i,j}$, we define

$\mathcal{C}_{i,j}^o(x) \in \{0, 1\}$ to indicate whether x exhibits an *objective conflict* of this type. For instance, $\mathcal{C}_{VP}^o(x) = 1$ indicates that X_V contradicts the parametric prior for a specific fact. This is an input-level property and does not itself imply the model will engage with the conflict during inference, motivating a process-level notion.

Effective Conflict characterizes whether an objective conflict is triggered during reasoning and reflected in the internal state: $\mathcal{C}_{i,j}^e(t \mid x) \in \{0, 1\}$ indicates whether, at step t , the model relies on mutually incompatible information of type $\mathcal{C}_{i,j}$. The relationship is asymmetric:

$$\mathbb{P}(\mathcal{C}_{i,j}^e(t \mid x) = 1 \mid \mathcal{C}_{i,j}^o(x) = 1) < 1, \quad (2)$$

i.e., *objective conflict* captures whether a conflict exists at the input level (jointly induced by input and priors), while *effective conflict* captures whether and when it activates during reasoning (model- and process-dependent).

Objective conflict data construction. For mechanistic analysis, we construct a benchmark of isolated pairwise conflicts, with each example containing one type (VP, VT, or PT). This serves as a diagnostic stress-test of conflict arbitration under contradiction, not an estimate of in-the-wild prevalence. For each x , we generate a long-CoT trajectory and align input-level labels $\mathcal{C}_{i,j}^o(x)$ with step-level *effective conflict* signals $\{\mathcal{C}_{i,j}^e(t \mid x)\}_{t=1}^T$ inferred from outputs. Table 1 reports activation statistics. Details are in Appendix A.

4 Probing Conflict from Internal States

In Section 3, we formalize knowledge conflict as input-level $\mathcal{C}_{i,j}^o(x)$ and process-level $\mathcal{C}_{i,j}^e(t \mid x)$. This section addresses the core question: *Is $\mathcal{C}_{i,j}^e(t \mid x)$ reflected in the model’s internal states, and can it be identified online during streaming generation?*

4.1 Token-level Probing of Conflict

We construct a streaming detector that, at the t -th token, determines whether an effective conflict $\mathcal{C}_{i,j}^e(t \mid x)$ is triggered from the hidden state $\mathbf{h}_t^{(l)}$ using only the prefix $y_{<t}$. Extending binary hallucination probes (Obeso et al., 2025), we cast detection as four-class classification: $z = 0$ denotes no conflict ($\mathcal{C}_{i,j}^e(t \mid x) = 0, \forall(i, j)$), and $z \in \{1, 2, 3\}$ correspond to active $\mathcal{C}_{VP}$, $\mathcal{C}_{PT}$, and $\mathcal{C}_{VT}$.

Formally, we define a probe f_ϕ that maps hidden states to conflict-label probabilities:

$$P_\phi(z \mid \mathbf{h}_t^{(l)}) = \text{Softmax}\left(f_\phi(\mathbf{h}_t^{(l)})\right), \quad (3)$$

$$z \in \{0, 1, 2, 3\}.$$

Supervision comes from span-level assertion annotations in Table 1; we project each span label to its tokens to obtain the dense label sequence $\{z_t\}$. Since conflict tokens are extremely sparse in long-CoT, we train f_ϕ with a weighted cross-entropy:

$$\mathcal{L}_{\text{probe}} = - \sum_t w_t \log P_\phi(z_t \mid \mathbf{h}_t^{(l)}), \quad (4)$$

where w_t upweights $z \in \{1, 2, 3\}$ to prevent collapse onto the no-conflict majority class. Training details, stricter prefix-only protocols, surface-signal control baselines, and a blind human check of span supervision are reported in Appendix C.

4.2 Verifying the Separability of Conflicts

We evaluate whether the probe reliably diagnoses conflicts from internal states, examining token-level separability of *effective conflicts* and sample-level recovery of *objective conflict* types.

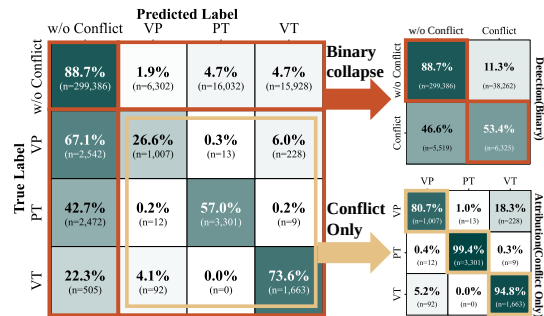


Figure 2: **Token-level separability of effective conflict** $\mathcal{C}_{i,j}^e(t \mid x)$. The left panel shows the confusion matrix over token-level conflict predictions. The right panels decompose performance into binary detection of conflict versus no-conflict, and fine-grained attribution among conflict types. Values denote row-normalized recall.

(I) Separability of Effective Conflicts: Local Signals in Sparse Regimes. We first ask whether the probe can distinguish different types of *effective conflicts* $\mathcal{C}_{i,j}^e(t \mid x)$ from internal states. As shown in Figure 2, the probe demonstrates robust discrimination. In binary detection (Stage I), it achieves a high True Negative rate of 88.7%, effectively filtering non-conflicting steps; the 46.6% False Negative rate is primarily driven by semantic sparsity within conflict spans (67.1% of $\mathcal{C}_{VP}$ tokens are misclassified due to weak local signals). Once *effective conflict* is activated (Stage II), separability between conflict types sharply increases: $\mathcal{C}_{PT}$ and $\mathcal{C}_{VT}$ reach 99.4% and 94.8%; even the subtlest $\mathcal{C}_{VP}$ jumps from 26.6% (global) to 80.7% (conditioned). The minimal off-diagonal confusion ($< 1\%$ between PT and others) confirms that distinct conflict types occupy highly separable subspaces.

Because the three types are built from different corpora, separability could in principle reflect dataset or template identity. Our contrast is *within-trajectory*: conflict and no-conflict tokens of the same sample share the same image, prompt, and template, so any input- or corpus-level confound is constant within a sample and cannot produce this separation. Consistent with this, the probe fires intermittently rather than uniformly (token density 1.65–11.77%, Table 1), surface cues (position, perplexity) stay far weaker under the same protocol (≤ 68.5 vs. 93.4–98.4 AUC; Appendix C.10), and detection is prefix-only (first-token 89.8%, window $w=8$ 95.1% AUC; Appendix C.9).

Conclusion (Local Effective Conflicts): Different *effective conflict* types yield distinct local structure in the internal states, recoverable in a streaming fashion under strict prefix-only protocols, although per-type detectability varies with the semantic density of each conflict.

(II) Alignment to Objective Conflicts: Aggregating Effective Signals. We next examine whether aggregating local *effective conflicts* $\mathcal{C}_{i,j}^e(t \mid x)$ along a reasoning trajectory recovers the corresponding *objective conflict* $\mathcal{C}_{i,j}^o(x)$, testing robustness beyond individual steps.

For each long-CoT trajectory, we mean-pool hidden states of activated *effective conflicts* into a sample-level representation. The t-SNE visualization in Figure 3 shows samples sharing the same *objective conflict* type form compact, well-separated clusters, indicating consistent global structure. Quantitatively, we infer the *objective*

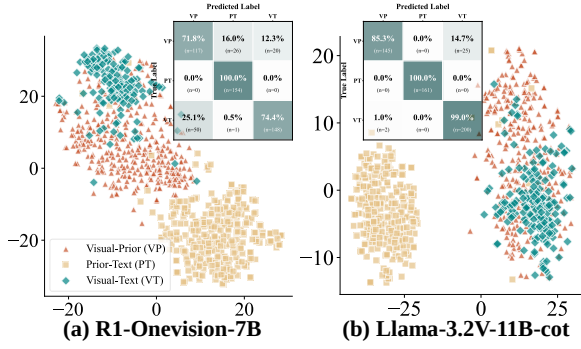


Figure 3: **Sample-level separability of conflict types.** We visualize the t-SNE projection of hidden states at layer 20 (R1-Onevision) and layer 39 (Llama-3.2V). The three conflict categories are colored by their **Objective Conflict** labels, defined during dataset construction. The top-right confusion matrices illustrate sample-level attribution performance.

conflict type via

$$\hat{\mathcal{C}}_{\text{sample}} = \arg \max_{\mathcal{C}_{i,j}} \sum_{t=1}^T \mathbb{I}[\mathcal{C}_{i,j}^e(t | x) = 1]. \quad (5)$$

Comparing $\hat{\mathcal{C}}_{\text{sample}}$ with the ground-truth $\mathcal{C}_{i,j}^o(x)$ directly tests whether the model’s internal conflict aligns with the input-level conflict structure.

This aggregation uses the probe’s per-token predictions; *no* span annotation is available at test time. Annotated span boundaries are used to report position-resolved detection metrics, i.e., to evaluate *where* detection fires, not for the detector to run.

As shown in the inset matrices of Figure 3, aggregation substantially enhances separability. $\mathcal{C}_{\text{PT}}$ reaches a perfect 100.0% on both R1-Onevision and Llama-3.2V, confirming text–prior conflicts induce uniquely stable internal shifts. Residual confusion is largely confined to visual-conflict types (e.g., 25.1% of $\mathcal{C}_{\text{VT}}$ mistaken as $\mathcal{C}_{\text{VP}}$ on R1-Onevision; 14.7% of $\mathcal{C}_{\text{VP}}$ as $\mathcal{C}_{\text{VT}}$ on Llama-3.2V). This error structure is semantically organized rather than arbitrary, since both visually grounded types share a common visual-grounding component. We therefore read the shared VP/VT boundary as an empirical finding rather than as noise. Without re-training, the same probe transfers to five external benchmarks with different datasets and templates (binary AUC 87.15%; Appendix C.11), indicating that the recovered structure is not specific to our construction, though it does not establish reliable attribution under mixed or compositional conflicts.

Conclusion (Global Effective Conflicts): Aggregating stepwise *effective conflict* signals along the reasoning trajectory makes *objective conflict* types $\mathcal{C}_{i,j}^o(x)$ robustly separable at the **sample level**: *effective conflicts* form consistent global patterns aligned with the input-level conflict structure rather than acting only as local artifacts.

4.3 Cross-Layer Conflict Signals

We scan model depth to localize where *effective* conflicts are most strongly encoded. For each layer l , we train the same token-level probe on $\mathbf{h}_t^{(l)}$ and report one-vs-rest AUC / Recall@0.1 for $\{\mathcal{C}_{\text{VP}}, \mathcal{C}_{\text{PT}}, \mathcal{C}_{\text{VT}}\}$.

Beyond probe separability, we also quantify a lightweight mechanistic correlate (Huang et al., 2025a): how attention-head activations differ between *conflict* and *no-conflict* token positions. Let $\mathcal{A}^{(l)}$ denote attention heads at layer l and $\mathbf{o}_t^{(l,a)}$ the output of head a at token t . Let $\mathcal{S}_{\text{conf}}$ collect the (x, t) with $\mathcal{C}_{i,j}^e(t | x) = 1$ for some (i, j) , and $\mathcal{S}_{\text{nconf}}$ those with no active conflict. The layer-wise activation ratio on a token set $\mathcal{S}$ is

$$R^{(l)}(\mathcal{S}) = \mathbb{E}_{(x,t) \in \mathcal{S}} \frac{1}{|\mathcal{A}^{(l)}|} \sum_{a \in \mathcal{A}^{(l)}} \mathbb{I}[\|\mathbf{o}_t^{(l,a)}\|_2 > \gamma], \quad (6)$$

with γ a fixed activation threshold (Appendix C.3), and the activation drift $\Delta R^{(l)} = R^{(l)}(\mathcal{S}_{\text{conf}}) - R^{(l)}(\mathcal{S}_{\text{nconf}})$ measures how strongly attention shifts when effective conflicts are triggered.

Figure 4 reveals distinct depth-dependent signatures from both measurements. **(I) Probe Separability:** In 7B models (R1-Onevision, Ocean-R1), discrimination rises in early layers and peaks in the mid-to-late block (Layers 15–22), where AUC for $\mathcal{C}_{\text{PT}}$ and $\mathcal{C}_{\text{VT}}$ consistently exceeds 93% before declining. Llama-3.2V pushes saturation deeper, maintaining $\geq 95\%$ separability up to Layer 39. **(II) Activation Drift** aligns with these attention shifts: R1-series shows negative drift (suppression) peaking at Layers 18–22, while Llama-3.2V shows positive drift (enhancement) in Layers 30–39. We refer to this co-located region as a *conflict-sensitive layer band* and probe at Layer 20 for 7B and Layer 39 for 11B. The band is broad rather than a single privileged layer. The precise location varies with architecture and training, but in all three backbones the signal is depth-localized rather than uniformly distributed (Appendix C.7).

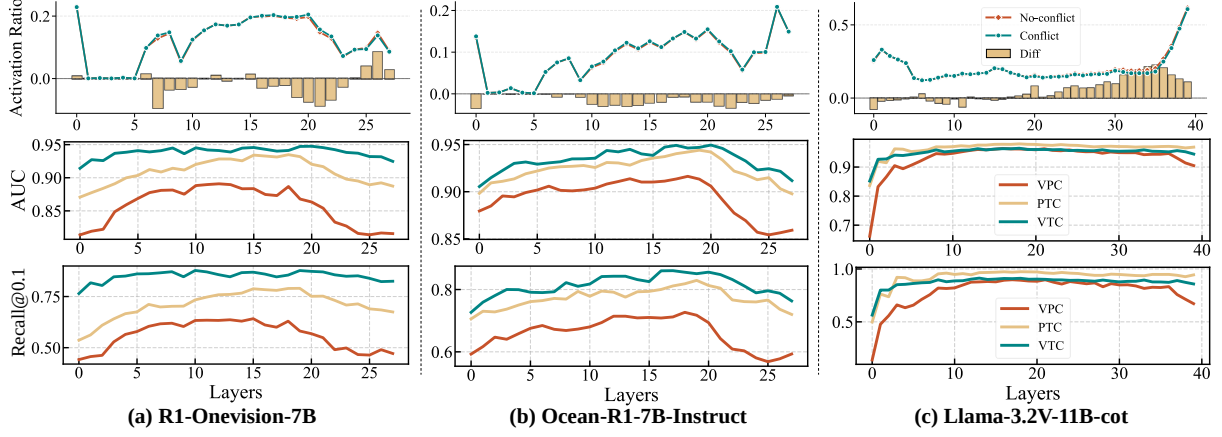


Figure 4: **Cross-layer conflict signals.** Top row: attention-head activation ratio on *conflict* tokens vs. *no-conflict* tokens (lines), and their difference (bars), computed using effective conflict labels. Middle/bottom rows: layer-wise probe performance (one-vs-rest AUC and Recall@0.1) for $\mathcal{C}_{VP}$, $\mathcal{C}_{PT}$, $\mathcal{C}_{VT}$ across three MLLM backbones.

Conclusion (Layer-level): Probe separability and attention drift co-localize in a backbone-specific mid-to-late layer band, indicating that conflict-related signals are **depth-localized** rather than uniformly distributed, even though the exact band depends on the model.

Conclusion (Linearity): A linear probe matches the non-linear MLP, indicating *effective conflicts* are explicitly **linearly separable** rather than entangled in non-linear manifolds, enabling efficient inference-time detection.

4.4 Linearity of Conflict Representation

To assess whether *effective conflict* signals $\mathcal{C}_{i,j}^e(t | x)$ encoded in $\mathbf{h}_t^{(l)}$ are explicitly linear or non-linearly entangled, we run experiments at the conflict-sensitive band layers identified in Section 4.3, using two probe architectures with distinct assumptions: **(I) Linear Probe (f_{lin}):** a single projection $\mathbf{W} \in \mathbb{R}^{d \times 4}$ evaluating *linear separability*, where high accuracy indicates clear, decoupled conflict boundaries. **(II) MLP Probe (f_{mlp}),** a deep MLP with three dimension-reducing layers ($1024 \rightarrow 512 \rightarrow 256$) and ReLU, designed to capture *non-linear entanglement*.

As shown in Table 2, we report AUC and Recall@0.1 under “Span-Max” aggregation (max predicted probability across tokens within each span; Appendix C.5). The Linear Probe achieves strong performance across all conflict types (AUC 93.2–98.8%, Recall@0.1 76.9–97.8%). For $\mathcal{C}_{PT}$, Linear achieves 98.6% AUC and 96.1–97.2% Recall@0.1; for $\mathcal{C}_{VP}$ and $\mathcal{C}_{VT}$, 93.4–95.9% AUC and 76.9–87.9% Recall@0.1, comparable to MLP. That a single linear layer suffices indicates that for knowledge conflicts, LLM features are already explicitly disentangled in the high-dimensional space; non-linear complexity yields no significant gain.

5 Intervening in Knowledge Conflict

Building on Section 4’s finding that *effective conflicts* $\mathcal{C}_{i,j}^e(t | x)$ are streaming-decodable and linearly separable in mid-to-late layers, we ask: *given an input with $\mathcal{C}_{i,j}^o(x) = 1$, can inference-time interventions bias the model toward a desired knowledge source, or suppress effective conflict activation during generation?*

5.1 Unified directional interventions

Two control objectives. We consider two settings under objectively conflicting inputs. **(I) Targeted source control.** Given a target source $\mathcal{K}_s \in \{\mathcal{K}_i, \mathcal{K}_j\}$, we intervene so the model follows $\mathcal{K}_s$, yielding two directions: **Forward** (toward the truth-anchored, benchmark-reliable source) and **Reverse** (toward the opposite source). **(II) Conflict mitigation.** We measure whether interventions reduce *effective conflict* activation, using the expected fraction of conflict-detected reasoning steps:

$$\mathbb{E}_x \mathbb{E}_t [\mathbb{I}[\mathcal{C}_{i,j}^e(t | x) = 1]]. \quad (7)$$

A unified view of directional interventions. Let $\ell_t \in \mathbb{R}^{|\mathcal{V}|}$ denote the pre-softmax logits at step t . We view an inference-time intervention as an additive logit perturbation, applied directly or implicitly via hidden-state manipulation:

$$\tilde{\ell}_t = \text{softmax}(\ell_t + \Delta\ell_t), \quad \Delta\ell_t = \mathcal{I}(x, y_{<t}). \quad (8)$$

Table 2: **Assessment of conflict probe performance across three VLM backbones.** We report AUC and Recall at FPR=0.1 (Rec@0.1) under the One-vs-Rest setting. Gray rows indicate the **Span-Max** aggregation, which consistently outperforms token-level baselines. Values are presented as percentages (%).

Models	Probe	Granularity	AUC (%)				Recall@0.1 (%)			
			w/o Conflict	$\mathcal{C}_{VP}^o = 1$	$\mathcal{C}_{PT}^o = 1$	$\mathcal{C}_{VT}^o = 1$	w/o Conflict	$\mathcal{C}_{VP}^o = 1$	$\mathcal{C}_{PT}^o = 1$	$\mathcal{C}_{VT}^o = 1$
R1-Onevision (7B)	Linear	All Token	81.7 $\pm$ 0.1	86.3 $\pm$ 0.2	92.0 $\pm$ 0.1	94.8 $\pm$ 0.2	50.0 $\pm$ 0.3	56.8 $\pm$ 0.2	75.1 $\pm$ 0.1	87.3 $\pm$ 0.3
		Span Only	76.8 $\pm$ 0.2	82.5 $\pm$ 0.1	90.8 $\pm$ 0.3	95.4 $\pm$ 0.2	35.5 $\pm$ 0.2	44.5 $\pm$ 0.3	70.5 $\pm$ 0.2	88.5 $\pm$ 0.1
		Span-Max	93.2$\pm$0.1	94.2$\pm$0.2	98.6$\pm$0.1	97.3$\pm$0.1	81.5$\pm$0.2	82.4$\pm$0.1	97.2$\pm$0.1	93.8$\pm$0.2
	MLP	All Token	95.5 $\pm$ 0.1	90.4 $\pm$ 0.2	85.2 $\pm$ 0.3	94.1 $\pm$ 0.1	89.1 $\pm$ 0.2	68.4 $\pm$ 0.1	62.7 $\pm$ 0.2	79.3 $\pm$ 0.1
		Span Only	95.7 $\pm$ 0.2	86.1 $\pm$ 0.3	80.3 $\pm$ 0.1	93.3 $\pm$ 0.2	89.8 $\pm$ 0.1	53.0 $\pm$ 0.2	43.7 $\pm$ 0.2	76.8 $\pm$ 0.3
		Span-Max	97.3$\pm$0.1	94.5$\pm$0.1	93.2$\pm$0.2	99.1$\pm$0.1	93.4$\pm$0.2	82.4$\pm$0.1	82.1$\pm$0.1	98.7$\pm$0.2
Ocean-R1 (7B-Instruct)	Linear	All Token	83.0 $\pm$ 0.2	90.6 $\pm$ 0.1	94.2 $\pm$ 0.2	94.9 $\pm$ 0.1	53.7 $\pm$ 0.3	69.4 $\pm$ 0.1	81.3 $\pm$ 0.2	85.6 $\pm$ 0.1
		Span Only	78.5 $\pm$ 0.1	86.7 $\pm$ 0.3	90.0 $\pm$ 0.2	97.6 $\pm$ 0.1	41.4 $\pm$ 0.2	52.5 $\pm$ 0.2	66.6 $\pm$ 0.1	94.6 $\pm$ 0.3
		Span-Max	95.0$\pm$0.2	95.9$\pm$0.1	98.6$\pm$0.1	98.8$\pm$0.1	85.7$\pm$0.1	87.9$\pm$0.2	97.1$\pm$0.1	97.8$\pm$0.2
	MLP	All Token	95.5 $\pm$ 0.1	92.8 $\pm$ 0.1	85.0 $\pm$ 0.2	95.5 $\pm$ 0.1	87.1 $\pm$ 0.2	75.6 $\pm$ 0.3	61.6 $\pm$ 0.1	85.2 $\pm$ 0.2
		Span Only	97.8 $\pm$ 0.2	87.3 $\pm$ 0.2	79.7 $\pm$ 0.1	91.7 $\pm$ 0.2	95.7 $\pm$ 0.3	53.9 $\pm$ 0.1	43.3 $\pm$ 0.2	71.0 $\pm$ 0.1
		Span-Max	99.2$\pm$0.1	96.5$\pm$0.1	95.3$\pm$0.2	98.4$\pm$0.1	98.9$\pm$0.1	89.8$\pm$0.2	87.5$\pm$0.1	96.1$\pm$0.1
Llama-3.2V (11B-coT)	Linear	All Token	88.7 $\pm$ 0.2	90.5 $\pm$ 0.1	96.9 $\pm$ 0.2	94.5 $\pm$ 0.1	68.4 $\pm$ 0.3	67.2 $\pm$ 0.2	94.4 $\pm$ 0.1	85.8 $\pm$ 0.2
		Span Only	79.6 $\pm$ 0.2	85.8 $\pm$ 0.2	90.2 $\pm$ 0.1	95.2 $\pm$ 0.3	43.2 $\pm$ 0.1	51.1 $\pm$ 0.2	66.0 $\pm$ 0.2	88.4 $\pm$ 0.1
		Span-Max	93.9$\pm$0.1	93.4$\pm$0.1	98.4$\pm$0.2	97.2$\pm$0.1	83.5$\pm$0.2	76.9$\pm$0.1	96.1$\pm$0.2	93.1$\pm$0.1
	MLP	All Token	95.8 $\pm$ 0.2	90.7 $\pm$ 0.1	88.7 $\pm$ 0.2	96.9 $\pm$ 0.1	89.4 $\pm$ 0.1	64.3 $\pm$ 0.2	70.6 $\pm$ 0.3	93.7 $\pm$ 0.2
		Span Only	96.1 $\pm$ 0.1	85.5 $\pm$ 0.3	79.2 $\pm$ 0.2	89.2 $\pm$ 0.1	90.8 $\pm$ 0.2	46.7 $\pm$ 0.1	40.5 $\pm$ 0.2	65.2 $\pm$ 0.1
		Span-Max	97.2$\pm$0.1	94.5$\pm$0.2	93.4$\pm$0.1	97.8$\pm$0.1	93.2$\pm$0.1	82.3$\pm$0.2	82.3$\pm$0.1	94.4$\pm$0.2

We consider three instantiations of $\mathcal{I}$, ordered by the granularity at which they use the conflict signal: **(I) Visual contrastive decoding (VCD)** (Leng et al., 2024), a logit-level correction restricted to visual conflicts. **(II) Representation steering.** Leveraging the linear separability from Section 4, we adopt steering (Zhang et al., 2025c) that shifts the hidden state at a conflict-sensitive layer by a learned direction, $\tilde{\mathbf{h}}_t = \mathbf{h}_t + \lambda \mathbf{v}$ (λ : strength, $\mathbf{v}$: direction). **(III) Probe-guided control.** We use the streaming probe to score candidate continuations, reweighting decoding toward options less likely to trigger conflicts. For top- k candidates $\mathcal{V}_k$ with base probabilities $p_t(w)$:

$$\tilde{p}_t(w) \propto p_t(w) \exp(\alpha P_t^{(w)}), \quad w \in \mathcal{V}_k, \quad (9)$$

where $P_t^{(w)}$ is the probe-predicted no-conflict probability for continuation w , and α controls guidance strength (implementation details: Appendices D.4 and D.5).

5.2 Targeted source control

We evaluate whether targeted interventions bias the model toward a specified knowledge source under objectively conflicting inputs. We use an automated assertion-level judge, implemented with a strong off-the-shelf LLM, to assess semantic alignment with the target source: the judge extracts factual claims from the output and verifies each against the

corresponding truth anchor (image, input text, or world knowledge), yielding compact metrics **ASR** (Anchor Support Rate), **ARR** (Anchor Refutation Rate), and **OER** (Obvious Error Rate). Human verification on a stratified 10% subset ($\sim 1,500$ spans) yielded high agreement ($\kappa = 0.87$), confirming close alignment with human perception (Appendix D.2); a separate blind human evaluation of intervention outputs (Appendix D.8) gives the same directional pattern as the judge.

As shown in Figure 5, targeted source control is feasible but directionally asymmetric. **Forward** interventions (toward the truth-anchored source; vision for VP/VT and prior for PT) reliably improve semantic alignment, whereas **Reverse** control (forcing the competing source) often degrades it. We attribute this to an internal source-reliability prior: its persistence across three distinct backbones and a source-balanced protocol that decouples direction from source identity (Appendix D.9) suggests shared instruction-tuning biases toward user-provided context (Sharma et al., 2024; Zhang et al., 2025c), rather than construction artifacts.

Under **Forward** control, **Probe-guided control** improves ASR while lowering OER by $\sim 30\%$; **VCD** yields stronger but selective gains on $\mathcal{C}_{VP}$ (ASR +15%, ARR halved). **Reverse** control remains challenging: most methods regress or show negligible gains. Functionally, the probe mainly

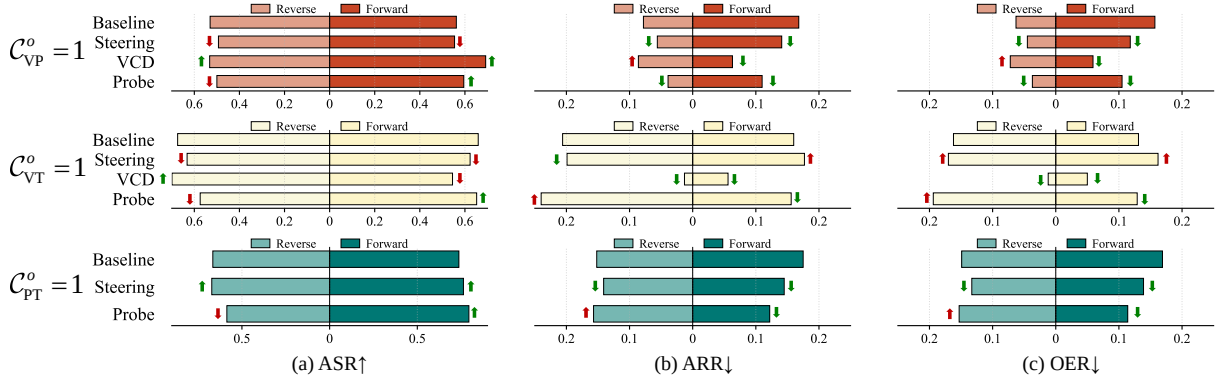


Figure 5: **Semantic performance of targeted source control.** We evaluate three conflict subsets ($\mathcal{C}_{VP}^o, \mathcal{C}_{VT}^o, \mathcal{C}_{PT}^o$) using judge-based metrics: **ASR** (Anchor Support Rate, $\uparrow$), **ARR** (Anchor Refutation Rate, $\downarrow$), and **OER** (Obvious Error Rate, $\downarrow$). **Forward/Reverse** denote interventions toward the truth-anchored (benchmark-reliable) vs. conflicting source. Arrows show changes relative to baseline. VCD is inapplicable to the non-visual $\mathcal{C}_{PT}^o$ subset.

Table 3: **Probe-independent source-anchored evaluation** (Ocean-R1). Paired differences (Probe-guided – Baseline, points) with 95% bootstrap CIs over matched question IDs ($n=40/\text{subset}$); **bold** marks intervals excluding zero. Per-arm scores: Appendix D.7.

Subset	$\Delta\text{ASR}\uparrow$	$\Delta\text{ARR}\downarrow$	$\Delta\text{OER}\downarrow$
$\mathcal{C}_{VP}^o=1$	0.0 [−3.8, 4.1]	+2.3 [−0.9, 5.4]	+2.3 [−0.9, 5.3]
$\mathcal{C}_{PT}^o=1$	+1.9 [−5.9, 10.4]	−6.2 [−12.7, −0.5]	−6.0 [−12.5, −0.5]
$\mathcal{C}_{VT}^o=1$	+21.0 [12.2, 30.2]	−15.1 [−22.8, −7.5]	−15.1 [−23.0, −7.5]

suppresses conflict states rather than enforcing weaker-source selection: **VCD** is high-gain but direction-sensitive, while **Representation steering** reduces ARR/OER but rarely sustains ASR gains.

Probe-independent semantic evidence. Since the probe both guides one intervention and supplies the token-level metrics of Section 5.3, we add a *source-anchored* evaluation whose judge sees only the input, the reliable source, and the response, never hidden states or probe scores (Appendix D.7).

The effects are *heterogeneous*, and we do not claim uniform gains: they concentrate on $\mathcal{C}_{VT}^o$ (+21.0 ASR, −15.1 ARR/OER), while $\mathcal{C}_{PT}^o$ shows reliable ARR/OER reductions with an ASR change spanning zero, and $\mathcal{C}_{VP}^o$ leaves no headroom at 92.0% baseline ASR. A blind human study agrees directionally, preferring Probe-guided outputs in 80.0% of pairs for target-source compliance and 73.3% for factual fidelity ($N=15$; Appendix D.8).

Ruling out a weak reverse operator. The same steering operator on the same $\mathcal{C}_{VT}^o$ examples raises ASR by +20.2 [11.2, 29.6] Forward but only +5.8 [−5.9, 16.8] Reverse, an interval spanning zero,

for a gap of +14.4 [5.5, 24.0] (Appendix D.9). The asymmetry thus survives a matched-operator, matched-example test, though we claim it only *under the tested controls*.

Conclusion (Targeted Source Control).

Inference-time interventions exhibit a clear **directional asymmetry**: biasing the model toward truth-anchored sources is far easier than forcing the opposite, indicating a stable source-dependent response tendency that can be strengthened but is difficult to reverse under the operators tested here.

5.3 Default-direction conflict mitigation

Complementing the semantic evaluation in Section 5.2, we ask a process-level question: *under the default (Forward) direction, can we reduce the activation of effective conflicts $\mathcal{C}_{i,j}^e(t | x)$ during generation?* We use four token-level mitigation metrics: Support Score (SS), Conflict Rate (CR), Confidence-Adjusted Conflict (CAC), and Conflict Confidence Index (CCI).

Table 4 shows that **Probe-guided control** consistently suppresses conflict dynamics across backbones. On visually involved subsets ($\mathcal{C}_{VT}^o$), CR drops significantly (e.g., 0.10→0.02 on R1-Onevision); even when CR remains stable ($\mathcal{C}_{PT}^o$), confidence-aware measures reveal deeper suppression (CCI drops by 25%), so the intervention mitigates conflict *intensity* rather than only frequency. In contrast, rigid **Representation steering** and unguided **VCD** struggle to generalize; VCD even exacerbates conflict rates fivefold on Llama-3.2V for $\mathcal{C}_{VT}^o$ (0.03→0.15). A sample-level paired analysis (Appendix D.6) confirms Probe-guided control wins on $\geq 76.7\%$ of matched

Table 4: **Token-level conflict mitigation under the forward direction.** Results are reported on three objective-conflict subsets ($\mathcal{C}_{VP}^o = 1$, $\mathcal{C}_{VT}^o = 1$, and $\mathcal{C}_{PT}^o = 1$) across three backbones. We report four token-level mitigation metrics: **SS** $\uparrow$, **CR** $\downarrow$, **CAC** $\downarrow$ and **CCI** $\downarrow$ (metric definitions in Appendix D.3). VCD is not applicable when $\mathcal{C}_{PT}^o = 1$ and is therefore reported only for the first two subsets.

	Subset	R1-Onevision-7B				Ocean-R1-7B-Instruct				Llama-3.2V-11B-cot			
		SS $\uparrow$	CAC $\downarrow$	CCI $\downarrow$	CR $\downarrow$	SS $\uparrow$	CAC $\downarrow$	CCI $\downarrow$	CR $\downarrow$	SS $\uparrow$	CAC $\downarrow$	CCI $\downarrow$	CR $\downarrow$
baseline	$\mathcal{C}_{VP}^o = 1$	0.94	0.04	0.70	0.03	0.89	0.07	0.71	0.06	0.94	0.04	0.45	0.02
	$\mathcal{C}_{VT}^o = 1$	0.88	0.08	0.80	0.10	0.87	0.09	0.79	0.10	0.90	0.06	0.72	0.03
	$\mathcal{C}_{PT}^o = 1$	0.82	0.12	0.80	0.15	0.82	0.12	0.80	0.15	0.84	0.11	0.70	0.11
VCD	$\mathcal{C}_{VP}^o = 1$	0.92 $^{-0.02}$	0.05 $^{+0.01}$	0.69 $^{-0.00}$	0.04 $^{+0.01}$	0.90 $^{+0.01}$	0.06 $^{-0.01}$	0.69 $^{-0.01}$	0.05 $^{-0.01}$	0.85 $^{-0.09}$	0.08 $^{+0.04}$	0.63 $^{+0.18}$	0.06 $^{+0.05}$
	$\mathcal{C}_{VT}^o = 1$	0.90 $^{+0.01}$	0.07 $^{-0.01}$	0.79 $^{-0.01}$	0.08 $^{-0.01}$	0.92 $^{+0.05}$	0.05 $^{-0.03}$	0.75 $^{-0.04}$	0.05 $^{-0.05}$	0.78 $^{-0.12}$	0.12 $^{+0.06}$	0.69 $^{-0.03}$	0.15 $^{+0.11}$
steering	$\mathcal{C}_{VP}^o = 1$	0.92 $^{-0.02}$	0.05 $^{+0.01}$	0.69 $^{-0.00}$	0.05 $^{+0.02}$	0.89 $^{-0.01}$	0.07 $^{+0.00}$	0.71 $^{+0.00}$	0.07 $^{+0.01}$	0.92 $^{-0.02}$	0.05 $^{+0.01}$	0.55 $^{+0.10}$	0.03 $^{+0.02}$
	$\mathcal{C}_{VT}^o = 1$	0.91 $^{+0.02}$	0.06 $^{-0.02}$	0.76 $^{-0.03}$	0.07 $^{-0.03}$	0.91 $^{+0.04}$	0.06 $^{-0.03}$	0.77 $^{-0.03}$	0.06 $^{-0.04}$	0.90 $^{+0.00}$	0.06 $^{-0.00}$	0.67 $^{-0.04}$	0.04 $^{+0.01}$
	$\mathcal{C}_{PT}^o = 1$	0.77 $^{-0.06}$	0.16 $^{+0.04}$	0.79 $^{-0.01}$	0.20 $^{+0.05}$	0.82 $^{+0.00}$	0.12 $^{-0.00}$	0.80 $^{-0.00}$	0.15 $^{+0.00}$	0.84 $^{+0.00}$	0.11 $^{-0.00}$	0.69 $^{-0.01}$	0.12 $^{+0.01}$
probe	$\mathcal{C}_{VP}^o = 1$	0.95 $^{+0.01}$	0.03 $^{-0.01}$	0.67 $^{-0.03}$	0.02 $^{-0.01}$	0.92 $^{+0.03}$	0.05 $^{-0.02}$	0.66 $^{-0.04}$	0.03 $^{-0.03}$	0.94 $^{+0.01}$	0.04 $^{-0.00}$	0.39 $^{-0.06}$	0.02 $^{-0.00}$
	$\mathcal{C}_{VT}^o = 1$	0.94 $^{+0.06}$	0.04 $^{-0.04}$	0.64 $^{-0.16}$	0.02 $^{-0.07}$	0.93 $^{+0.06}$	0.04 $^{-0.04}$	0.72 $^{-0.07}$	0.04 $^{-0.06}$	0.92 $^{+0.02}$	0.05 $^{-0.01}$	0.67 $^{-0.05}$	0.03 $^{-0.00}$
	$\mathcal{C}_{PT}^o = 1$	0.78 $^{-0.04}$	0.10 $^{-0.02}$	0.60 $^{-0.20}$	0.15 $^{+0.01}$	0.87 $^{+0.04}$	0.08 $^{-0.04}$	0.72 $^{-0.08}$	0.09 $^{-0.06}$	0.87 $^{+0.04}$	0.08 $^{-0.03}$	0.63 $^{-0.07}$	0.10 $^{-0.01}$

samples across all four metrics. This pattern is consistent with the mechanistic picture in Section 4: because the probe operates on the conflict-encoding subspace identified by linear separability, it can locally re-weight candidate continuations that would otherwise trigger *effective conflicts*, whereas global steering and contrastive corrections lack the same per-token specificity.

Conclusion (Conflict Mitigation). Guiding the model toward the reliable source attenuates internal conflict dynamics, reducing the intensity and frequency of *effective conflict* $\mathcal{C}_{i,j}^e(t \mid x)$ states. These activations are not inherent to generation, but plastic internal states that can be suppressed.

6 Conclusion

In this work, we study failures in multimodal long-CoT reasoning from the perspective of knowledge conflict, rather than knowledge absence. By distinguishing *objective conflicts* from *effective conflicts* during reasoning, we show that many failures arise from how conflicting knowledge is resolved over time. We find that *effective conflicts* are encoded as explicit and linearly decodable signals, concentrated in a mid-to-late conflict-sensitive layer band. Leveraging these signals, we uncover a pronounced directional asymmetry: guiding the model toward its reliability-aligned source is substantially easier than forcing conflict resolution in the opposite direction, indicating a biased and path-dependent mechanism. Our evidence is representational and functional rather than circuit-level, and our benchmark isolates one source pair at a time. Looking forward, we hope this perspective motivates analysis and control methods for richer conflict structures

and more complex multimodal reasoning settings.

Limitations

Our benchmark deliberately isolates a single pairwise conflict type per example to enable controlled mechanistic analysis, so the reported activation statistics characterize MLLM behavior *under* conflict rather than the natural prevalence of such conflicts in deployment. While we provide cross-domain evidence on five external benchmarks (Appendix C.11) and control baselines that rule out shortcut learning (Appendix C.10), generalization to wild long-CoT trajectories with mixed, dynamic, or compositional conflicts remains only partially explored. Conflict span annotation and semantic evaluation both rely on strong off-the-shelf LLMs; we mitigate this with stratified human verification ($\kappa=0.87$, Appendix D.2), focused blind annotation studies (Appendix C.12), and blind human evaluation of intervention outputs (Appendix D.8), but cannot fully exclude residual judge bias given the limited scale of human studies.

Methodologically, probes are correlational tools that identify *where* information is linearly readable, not the causal circuits that compute it; our lightweight interventions (VCD, representation steering, probe-guided control) test functional relevance of the identified subspaces but do not establish full mechanistic understanding, which would require activation patching or circuit-level ablation. Experiments cover three open-weight backbones (R1-Onevision, Ocean-R1, Llama-3.2V) at two parameter scales; the *conflict encoding stage* is depth-localized in all three but its exact layer depends on architecture and training (Section 4.3).

Ethical Considerations

This work aims to advance the understanding and reliability of MLLMs in long-CoT reasoning by diagnosing how models arbitrate between visual, textual, and parametric knowledge sources, with the broader goal of reducing the propagation of hallucinated or factually inconsistent content. We acknowledge a dual-use risk: inference-time interventions that bias a model toward a target source could in principle be repurposed to *strengthen* a chosen source against truthful evidence. Our finding that *Forward* (truth-anchored) interventions are substantially easier than *Reverse* (anti-truth) ones (Section 5.2) limits but does not eliminate this risk; we therefore caution that probe-guided or steering-based methods should not be deployed in user-facing systems without first auditing the targeted source for factual reliability. The diagnostic benchmark is built on top of publicly available multimodal corpora used under their respective licenses; no personally identifiable information or sensitive demographic content is introduced. AI assistants were used only for language polishing, in accordance with the ACL Policy on AI Writing Assistance, and the authors take full responsibility for the correctness of the manuscript.

Acknowledgments

The work was supported by the National Key R&D Program of China under Grant 2025YFE0216500, the Major Program of National Natural Science Foundation of China under Grant 62595802, the Foundation for Outstanding Research Groups of Hubei Province of China under Grant 2025AFA012, and the 111 Project on Computational Intelligence and Intelligent Control under Grant B18024.

References

- Wenbin An, Jiahao Nie, Feng Tian, Haonan Lin, mingxiang cai, Yaqiang Wu, QianYing Wang, Xiaoqin Zhang, and Shijian Lu. 2025. Boosting knowledge utilization in multimodal large language models via adaptive logits fusion and attention reallocation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.
- Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618.
- Hanzhu Chen, Xu Shen, Jie Wang, Zehao Wang, Qitan Lv, Junjie He, Rong Wu, Feng Wu, and Jieping Ye. 2025. Knowledge graph finetuning enhances knowledge manipulation in large language models. In *The Thirteenth International Conference on Learning Representations*.
- Siyuan Cheng, Ningyu Zhang, Bozhong Tian, Xi Chen, Qingbin Liu, and Huajun Chen. 2024. Editing language model-based knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17835–17843.
- Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3867–3876.
- Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2025a. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*.
- Yuntao Du, Kailin Jiang, Zhi Gao, Chenrui Shi, Zilong Zheng, Siyuan Qi, and Qing Li. 2025b. MMKE-bench: A multimodal editing benchmark for diverse visual knowledge. In *The Thirteenth International Conference on Learning Representations*.
- Aaron Fanous, Jacob Goldberg, Ank Agarwal, Joanna Lin, Anson Zhou, Sonnet Xu, Vasiliki Bikia, Roxana Daneshjou, and Sanmi Koyejo. 2025. Syceval: Evaluating llm sycophancy. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 893–900.
- Jiahai Feng, Stuart Russell, and Jacob Steinhardt. 2025. Monitoring latent world states in language models with propositional probes. In *The Thirteenth International Conference on Learning Representations*.
- Aarti Ghatkesar, Uddeshya Upadhyay, and Ganesh Venkatesh. 2025. Looking beyond language priors: Enhancing visual comprehension and attention in multimodal models. *arXiv preprint arXiv:2505.05626*.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

- Yangyang Guo, Fangkai Jiao, Zhiqi Shen, Liqiang Nie, and Mohan Kankanhalli. 2024. Unk-vqa: A dataset and a probe into the abstention ability of multi-modal large models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. 2026. Learning to see before seeing: Demystifying LLM visual priors from language pre-training. In *The Fourteenth International Conference on Learning Representations*.
- Pengcheng Huang, Zhenghao Liu, Yukun Yan, Haiyan Zhao, Xiaoyuan Yi, Hao Chen, Zhiyuan Liu, Maosong Sun, Tong Xiao, Ge Yu, and 1 others. 2025a. Parammute: Suppressing knowledge-critical ffns for faithful retrieval-augmented generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Yukun Huang, Sanxing Chen, Hongyi Cai, and Bhuwan Dhingra. 2025b. To trust or not to trust? enhancing large language models’ situated faithfulness to external contexts. In *The Thirteenth International Conference on Learning Representations*.
- Nan Huo, Jinyang Li, Bowen Qin, Ge Qu, Xiaolong Li, Xiaodong Li, Chenhao Ma, and Reynold Cheng. 2025. Micro-Act: Mitigate knowledge conflict in question answering via actionable self-reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18550–18574. Association for Computational Linguistics.
- Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxian Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, and 1 others. 2025. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3(1):27.
- Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojuan Jiang, Kang Liu, and Jun Zhao. 2024. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.
- Jonathan Kahana, Eliahu Horwitz, Imri Shuval, and Yedid Hoshen. 2025. Deep linear probe generators for weight space learning. In *The Thirteenth International Conference on Learning Representations*.
- Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. 2024. Prismatic VLMs: Investigating the design space of visually-conditioned language models. In *Forty-first International Conference on Machine Learning*.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.
- Gaotang Li, Yuzhong Chen, and Hanghang Tong. 2025. Taming knowledge conflicts in language models. In *Forty-second International Conference on Machine Learning*.
- Kenneth Li, Samy Jelassi, Hugh Zhang, Sham M. Kakade, Martin Wattenberg, and David Brandfonbrener. 2024. Q-probe: A lightweight approach to reward maximization for language models. In *Forty-first International Conference on Machine Learning*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.
- Aiwei Liu, Sheng Guan, Yiming Liu, Leyi Pan, Yifei Zhang, Liancheng Fang, Lijie Wen, Philip S. Yu, and Xuming Hu. 2025a. Can watermarked LLMs be identified by users via crafted prompts? In *The Thirteenth International Conference on Learning Representations*.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.
- Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024b. Phd: A chatgpt-prompted visual hallucination evaluation dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jentsse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. 2025b. Insight over sight: Exploring the vision-knowledge conflicts in multimodal LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17825–17846. Association for Computational Linguistics.
- Yexin Liu, Zhengyang Liang, Yuezhe Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, and 1 others. 2025c. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9087–9097.
- Yifan Lu, Yigeng Zhou, Jing Li, Yequan Wang, Xuebo Liu, Daojing He, Fangming Liu, and Min Zhang. 2025. Knowledge editing with dynamic knowledge graphs for multi-hop question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24741–24749.

- Yujie Lu, Dongfu Jiang, Wenhui Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. 2024. Wild-vision: Evaluating vision-language models in the wild with human preferences. *Advances in Neural Information Processing Systems*, 37:48224–48255.
- Xinyu Ma, Ziyang Ding, Zhicong Luo, Chi Chen, Zonghao Guo, Derek F Wong, Xiaoyi Feng, and Maosong Sun. 2025. Deepception: Advancing r1-like cognitive visual perception in mllms for knowledge-intensive visual grounding. *arXiv preprint arXiv:2503.12797*.
- Lingfeng Ming, Yadong Li, Song Chen, Jianhua Xu, Zenan Zhou, and Weipeng Chen. 2025. Ocean-r1: An open and generalizable large vision-language model enhanced by reinforcement learning. Accessed: 2025-04-03.
- Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *The Thirteenth International Conference on Learning Representations*.
- Minheng Ni, YuTao Fan, Lei Zhang, and Wangmeng Zuo. 2025. Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. In *The Thirteenth International Conference on Learning Representations*.
- Oscar Obeso, Andy Ardit, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. 2025. Real-time detection of hallucinated entities in long-form generation. *arXiv preprint arXiv:2509.03531*.
- Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2024. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*.
- Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024. IRCAN: Mitigating knowledge conflicts in LLM generation via identifying and reweighting context-aware neurons. In *Advances in Neural Information Processing Systems*, volume 37, pages 4997–5024. Curran Associates, Inc.
- Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Chenmien Tan, Ge Zhang, and Jie Fu. 2024. Massive editing for large language models via meta learning. In *The Twelfth International Conference on Learning Representations*.
- Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.
- Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.
- Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, and 1 others. 2025a. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*.
- Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025b. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.
- Kevin Wu, Eric Wu, and James Zou. 2024. Clashes: Quantifying the tug-of-war between an llm’s internal prior and external evidence. *Advances in neural information processing systems*, 37:33402–33422.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.
- Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2087–2098.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

- Ziang Yan, Zhilin Li, Yinan He, Chenting Wang, Kun-chang Li, Xinhao Li, Xiangyu Zeng, Zilei Wang, Yali Wang, Yu Qiao, and 1 others. 2025. Task preference optimization: Improving multimodal large language models with vision task alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29880–29892.
- Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.
- Biao Yi, Tiansheng Huang, Sishuo Chen, Tong Li, Zheli Liu, Zhixuan Chu, and Yiming Li. 2025. Probe before you talk: Towards black-box defense against backdoor unalignment for large language models. In *The Thirteenth International Conference on Learning Representations*.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. RLhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.
- Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, and 1 others. 2025. RLaiF-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995.
- Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.
- Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mmllms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12401–12430.
- Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024b. How language model hallucinations can snowball. In *Forty-first International Conference on Machine Learning*.
- Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025a. FaithfulRAG: Fact-level conflict modeling for context-faithful retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21863–21882. Association for Computational Linguistics.
- Ruizhe Zhang, Yongxin Xu, Yuzhen Xiao, Runchuan Zhu, Xinke Jiang, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025b. Knowpo: Knowledge-aware preference optimization for controllable knowledge selection in retrieval-augmented language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25895–25903.
- Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, and 1 others. 2024c. Multitrust: A comprehensive benchmark towards trustworthy multimodal large language models. *Advances in Neural Information Processing Systems*, 37:49279–49383.
- Yu Zhang, Jinlong Ma, Yongshuai Hou, Xuefeng Bai, Kehai Chen, Yang Xiang, Jun Yu, and Min Zhang. 2025c. Evaluating and steering modality preferences in multimodal large language model. *arXiv preprint arXiv:2505.20977*.
- Zhang Zhang, Zhicheng Geng, Yuqing Ma, Tianbo Wang, Kai Lv, and Xianglong Liu. 2025d. Conflict-aware knowledge editing in the wild: Semantic-augmented graph representation for unstructured text. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Zhiqin Zhang, Yining Ma, Zhiguang Cao, and Hoong Chuin Lau. 2025e. Probing neural combinatorial optimization models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Zongmeng Zhang, Yufeng Shi, Jinhua Zhu, Wengang Zhou, Xiang Qi, peng zhang, and Houqiang Li. 2024d. Trustworthy alignment of retrieval-augmented large language models via reinforcement learning. In *Forty-first International Conference on Machine Learning*.
- Shiji Zhao, Ranjie Duan, Fengxiang Wang, Chi Chen, Caixin Kang, Shouwei Ruan, Jialing Tao, YueFeng Chen, Hui Xue, and Xingxing Wei. 2025. Jailbreaking multimodal large language models via shuffle inconsistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2045–2054.

A Dataset Construction and Annotation

A.1 Dataset Overview

We construct the **Objective Conflict Dataset**, a fine-grained benchmark designed to systematically evaluate knowledge conflict phenomena in MLLMs during long-CoT reasoning. Drawing upon the three-source knowledge framework established in Section 3.1, we leverage state-of-the-art closed-source models (e.g., GPT-5, Claude-Sonnet-4.5) to facilitate targeted instruction expansion, contradictory context generation, and span-level annotation. We emphasize that this dataset is a controlled stress-test benchmark designed for mechanistic analysis, rather than estimating conflict prevalence in natural deployments.

Role of Closed-Source Models and Truth Anchors. It is important to clarify that closed-source models serve exclusively as *construction and annotation tools*, rather than ground-truth generators. The truth anchors for conflict verification are derived from independent, objective sources:

- **Visual evidence** ($\mathcal{K}_{\text{vision}}$): The image X_V itself, which provides objectively verifiable visual facts independent of any model.
- **Textual input** ($\mathcal{K}_{\text{text}}$): The given prompt text X_T , which is explicitly provided and thus unambiguous.
- **Factual knowledge** ($\mathcal{K}_{\text{prior}}$): Human-annotated correct answers from established benchmarks (e.g., TruthfulQA (Lin et al., 2022)), ensuring that world knowledge verification relies on externally validated facts rather than model-generated content.

This design ensures that any potential biases in closed-source models affect only the *style* of generated queries and annotations, not the *correctness* of truth anchors. Furthermore, as we show in Appendix C.7, our core findings remain consistent across three trained backbones spanning two architecture families with different training recipes, suggesting that the observed phenomena are not artifacts of the annotation pipeline.

Our construction ensures that each sample contains a dominant pairwise objective conflict $\mathcal{C}_{i,j}^o(x) = 1$ (where $\mathcal{C}_{i,j} \in \{\mathcal{C}_{VP}, \mathcal{C}_{VT}, \mathcal{C}_{PT}\}$), while strictly minimizing cross-type contamination. The dataset covers the three core conflict types:

- **Vision-Text Conflict** ($\mathcal{C}_{VT}^o(x) = 1$): The objective visual evidence X_V contradicts the textual description X_T provided in the prompt.

- **Vision-Prior Conflict** ($\mathcal{C}_{VP}^o(x) = 1$): Counter-intuitive or synthetic visual inputs X_V contradict the parametric prior knowledge implicitly encoded in the model parameters θ .
- **Prior-Text Conflict** ($\mathcal{C}_{PT}^o(x) = 1$): Misleading textual instructions X_T contradict the parametric prior knowledge implicitly encoded in the model parameters θ .

This appendix details the pipeline used to construct these inputs and the subsequent annotation process used to identify effective conflict triggers $\mathcal{C}_{i,j}^e(t | x)$ during reasoning.

A.2 Data Sources and Selection

To ensure diversity and complexity, we curate samples from the **PhD** (Liu et al., 2024b) and **TruthfulQA** (Lin et al., 2022) benchmarks, adapting them to enforce specific objective conflict profiles.

Adaptation of PhD Benchmark. The PhD dataset, originally designed to evaluate visual faithfulness in MLLMs, spans tasks from object recognition to complex scene analysis. Two specific subsets are selected to construct visual-related conflicts. **Vision-Text Conflict.** The **PhD-icc** subset, containing 14,648 natural images paired with erroneous descriptions, is leveraged for this conflict type. Approximately 1,000 images are sampled from this collection to serve as ground-truth visual evidence X_V . We discard the original answers and contexts, instead employing advanced closed-source MLLMs to synthesize highly misleading textual contexts X_T and complex queries designed to cause conflict during reasoning. Then, we generate responses using the three open-source MLLMs under investigation to obtain their reasoning outputs under conflict. **Vision-Prior Conflict.** This study utilizes the **PhD-ccs** subset, which comprises 750 AI-generated images that deliberately violate physical laws or semantic common sense (e.g., a boat sailing on desert sand). In these scenarios, the visual evidence X_V inherently contradicts the model’s internal world knowledge $\mathcal{K}_{\text{prior}}$. We retain the core images and questions but discard the original answers. Each model performs independent inference, allowing us to analyze how they resolve the tension between perceptual evidence and parametric priors.

Adaptation of TruthfulQA Benchmark. For Prior-Text Conflict, we employ the **TruthfulQA** benchmark to construct conflicts where visual information is irrelevant or neutral. TruthfulQA

contains 817 questions across 38 categories (e.g., health, law, fiction) designed to elicit imitative falsehoods. These questions target scenarios where the model’s pre-training data might contain misconceptions. We utilize the adversarial “Best Incorrect Answers” provided by TruthfulQA to construct misleading textual contexts X_T that reinforce false beliefs, thereby creating a strong objective conflict against the factual truth.

A.3 Deceptive Context and Query Generation

To maximize the probability of triggering effective conflict $\mathcal{C}_{i,j}^e(t \mid x)$ during the reasoning process, we developed specialized strategies to synthesize queries. These queries are carefully synthesized to create high-intensity yet controlled contradictions, thereby reliably eliciting conflict-driven reasoning trajectories for analysis. The design principles for each conflict type are detailed below:

For $\mathcal{C}_{VT}$, we employ a Dual-Stage Prompting approach to synthesize high-intensity conflicts:

- **Step 1: Fact Extraction.** We first extract objective visual facts from the image X_V .
- **Step 2: Narrative Synthesis.** We utilize advanced closed-source LLMs to synthesize a coherent but factually contradictory narrative X_T , constructed to contradict key visual elements while maintaining superficial plausibility.
- **Step 3: Question Formulation.** We generate open-ended, non-binary questions that necessitate multi-step reasoning by fusing information from both modalities.

This strategy is designed to force the model into a conflict state, where it must explicitly choose to either generate a text-induced conflict or remain faithful to the visual evidence. The detailed prompt design principles are described in Section A.6.

For $\mathcal{C}_{VP}$, we append standardized constraints to the original queries to elicit longer reasoning trajectories. Specifically, we mandate the models to follow a structured four-step reasoning process, moving from an initial visual description and context analysis to an explicit evidence chain reasoning before reaching a final conclusion. By enforcing this structured Chain-of-Thought (CoT), we prevent the model from bypassing the conflict through brief or evasive answers. Instead, it compels the explicit manifestation of the clash between counter-intuitive visual evidence and internal priors within the generated text.

For $\mathcal{C}_{PT}$, we employ a structured template to embed the complete set of “Incorrect Answers” from

the TruthfulQA benchmark directly into the query. We present the original question followed by a deceptive instruction: “*Is it because of the following reasons? If not, please answer based on your knowledge.*” This design creates a direct conflict between the provided incorrect context X_T and the model’s parametric priors $\mathcal{K}_{\text{prior}}$. By explicitly listing multiple false justifications, we exert significant pressure on the model to conform to the provided misinformation. We then evaluate whether the model yields to these distractors or successfully retrieves the ground-truth answer from its internal knowledge.

A.4 Response Generation and Span-Level Annotation

To systematically observe how input level objective conflicts $\mathcal{C}_{i,j}^o(x)$ manifest as process-level effective conflicts $\mathcal{C}_{i,j}^e(t \mid x)$ during inference, we curate a targeted corpus of reasoning outputs generated by three representative MLLMs: R1-Onevision-7B, Ocean-R1-7B-Instruct, and Llama-3.2V-11B-cot. These outputs encompass a diverse range of model behaviors under conflicting inputs, providing the necessary empirical foundation for dissecting internal conflict dynamics.

Generation Settings. We perform inference using the vLLM framework with asynchronous batching. Key hyperparameters include: sampling temperature $T_{\text{temp}} = 0.7$ and maximum sequence length of 4,096 tokens, facilitating extended Chain-of-Thought (CoT) generation. For the Vision-Prior subset specifically, we employ a structured four-step reasoning process guided by the prompt design principles detailed in Section A.6.

Annotation Procedure. We implement an automated pipeline utilizing GPT-5 as a central judge. Adopting the *Extract-Verify-Score* paradigm, the process involves: (1) **Extraction** of model outputs into atomic, verifiable fact segments (*spans*). (2) **Verification** of each span against respective truth anchors ($\mathcal{K}$) and visual evidence X_V . and (3) **Scoring** by assigning conflict labels (e.g., *Supported*, $\mathcal{C}_{VP}^e$, $\mathcal{C}_{VT}^e$, or $\mathcal{C}_{PT}^e$) based on observed conflict-induced behaviors. Notably, to ensure precise localization, we employ a fuzzy string matching algorithm to align annotated spans back to their exact character-level indices within the original model output. This step corrects for minor textual rephras-

ing by the judge model, providing the granular supervision necessary for training our streaming conflict probes. The specific prompt used for annotation is shown in the box below.

A.5 Quality Control and Dataset Statistics

Human Verification. To ensure the reliability of the automated annotation process, we conduct human verification on a stratified random 10% subset of annotations (approximately 15,000 spans). Three annotators with graduate-level NLP background independently verified: (1) **span boundary accuracy**—whether the extracted span correctly delimits the atomic claim. (2) **conflict label correctness**—whether the assigned label ($\mathcal{C}_{VP}$, $\mathcal{C}_{VT}$, $\mathcal{C}_{PT}$, or *Supported*) accurately reflects the conflict type. Annotators achieved an inter-annotator agreement of $\kappa = 0.87$ (Cohen’s kappa) for conflict labels and $\kappa = 0.91$ for span boundaries. Disagreements were resolved through majority voting, with remaining ties adjudicated by a senior annotator.

The most common source of disagreement involved distinguishing $\mathcal{C}_{VP}$ from $\mathcal{C}_{VT}$ when both visual misinterpretation and textual influence were present. In such cases, annotators defaulted to the dominant conflict source based on explicit textual cues in the model’s reasoning chain.

Dataset Organization. The finalized dataset is organized into three model-specific subsets corresponding to R1-Onevision-7B, Ocean-R1-7B-Instruct, and Llama-3.2V-11B-cot. Within each subset, we apply stratified sampling to partition samples into Train (80%) and Test (20%) splits, ensuring balanced representation of the three conflict types ($\mathcal{C}_{VP}$, $\mathcal{C}_{VT}$, $\mathcal{C}_{PT}$) while minimizing distribution shift.

Final Statistics. The complete dataset comprises approximately 155,000 fine-grained annotated spans derived from model-specific reasoning trajectories. Detailed statistics regarding sample counts and span densities are provided in Table 5.

A.6 Prompt Design for Dataset Construction

This section details the prompt design principles used for dataset construction. We describe the key components of each prompt type and provide the complete annotation prompt used for span-level labeling.

Vision-Text Conflict Question Generation. To construct high-quality $\mathcal{C}_{VT}$ samples, we design a structured prompt that guides closed-source models through a five-stage process:

1. **Visual Fact Extraction:** The model first extracts 5–10 objective, verifiable visual facts from the image (e.g., number of people, visible objects, scene type, actions). These facts serve as ground-truth anchors for subsequent conflict verification.
2. **Conflicting Context Synthesis:** Based on extracted facts, the model generates a coherent but systematically misleading textual description (80–150 words). The key design principle is to create *plausible deception*: the text incorporates partially correct details (colors, approximate positions) while introducing at least three factual mismatches. For example, mislabeling leisure activities as emergencies while preserving neutral scene elements, or distorting object roles while maintaining spatial layout.
3. **Question Formulation:** The model designs open-ended, non-binary questions that necessitate multi-step reasoning by fusing both modalities. Questions are crafted to: (a) require at least 3 reasoning steps; (b) avoid explicitly indicating that the text is misleading; (c) encourage speculation about ambiguous image regions to amplify conflict probability.
4. **Reference Answer Generation:** An ideal, conflict-free answer is generated that relies solely on visual evidence, explicitly references extracted facts, and presents a clear multi-step reasoning chain.
5. **Conflict Tagging:** Each sample is tagged with the conflict type (`image_vs_text_conflict`) for downstream processing.

The output format is a structured JSON containing the conflicting context, question, and reference answer, enabling automated downstream processing.

Vision-Prior Conflict Reasoning Induction.

For $\mathcal{C}_{VP}$ samples, we employ a structured Chain-of-Thought induction prompt that mandates a four-step reasoning process:

1. **Visual Description:** Objectively describe visible elements without applying prior knowledge.
2. **Context Analysis:** Identify environmental settings and object relationships.
3. **Evidence Chain Reasoning:** Evaluate whether visual evidence aligns with physical laws or

Table 5: Statistics of the Objective Conflict Dataset and Generated Trajectories. The dataset provides span-level supervision for training conflict probes (Section 4) and evaluating interventions (Section 5).

Model Backbone	Split	Samples	Total Spans	Avg. Spans/Sample
R1-Onevision-7B	Test	498	9,038	18.1
	Train	1,987	34,300	17.3
Ocean-R1-7B	Test	495	12,881	26.0
	Train	1,978	51,931	26.3
Llama-3.2V-11B-cot	Test	513	9,419	18.4
	Train	2,051	37,118	18.1
Total	–	7,522	154,687	20.6

common sense, if conflict exists, explicitly articulate the tension.

4. **Conclusion:** Provide a final answer based on the preceding analysis.

This structured format prevents models from bypassing the conflict through brief or evasive answers, instead compelling explicit manifestation of the clash between counter-intuitive visual evidence and internalized priors.

Prior-Text Conflict Template. For $\mathcal{C}_{PT}$ samples, we embed “Best Incorrect Answers” from TruthfulQA directly into the query using a structured template: “[Original Question]. Is it because of the following reasons? [List of incorrect answers]. If not, please answer based on your knowledge.” This design creates direct pressure on the model to either conform to provided misinformation or retrieve factual knowledge from its parameters.

Span-Level Annotation Prompt. The annotation prompt implements our *Extract-Verify-Score* paradigm and is the most critical component for obtaining fine-grained supervision. The complete prompt is provided below.

Conflict Annotation Prompt

You are a multimodal fact-checking expert and conflict annotation engine.

Your task is to **EXHAUSTIVELY** fact-check a Vision-Language Model’s (VLM) completion against the given instruction, ground truth, and image, with a **PRIMARY GOAL** of **MAXIMUM SPAN COVERAGE** (high recall).

INPUTS

Instruction: <instruction>{instruction}</instruction>
Ground Truth: <groundtruth>{groundtruth}</groundtruth>

Model Completion: <completion>{completion}</completion>

CRITICAL TASK DEFINITION

You must identify and verify **ALL** minimal, self-contained, verifiable claims appearing in the completion.

A claim is **ANY** text span that asserts, implies, or assumes something that can be judged as true or false, including:

- existence of an object or entity
- object attributes (color, size, shape, material, state)
- quantities or counts
- spatial relations (left/right/on/in/near)
- actions or behaviors
- scene-level assumptions
- factual statements
- implicit assumptions derived from prior knowledge

IMPORTANT: Claims are **NOT** limited to named entities. Descriptive words, adjectives, numbers, directions, and verbs **OFTEN** contain conflicts. If a phrase can be checked, it **MUST** be extracted.

MANDATORY CLAIM DECOMPOSITION STEP

Before verification, you **MUST** mentally decompose the completion sentence by sentence. For **EACH** sentence:

- Identify **ALL** independent claims it contains.
- A single sentence usually contains **MULTIPLE** claims.
- Do **NOT** stop after extracting a few obvious entities.

WHAT TO VERIFY

You need to verify two types of content:

1. Entity-Level Information: People, organizations, locations, dates, statistics. Named or implied entities. Claims about multimodal content (e.g., “the building on the left”).

2. Visual Description Information: Object existence; Object attributes; Spatial relationships; Quantities; Actions and behaviors; Scene understanding.

SPAN EXTRACTION RULES (VERY STRICT)

For **EACH** verifiable claim:

1. Extract the **MINIMAL** possible span that directly expresses that claim.

2. The span **MUST** match the original text word-for-word.
3. Do **NOT** include surrounding context, articles, or extra description.
4. Do **NOT** merge multiple claims into one span.
5. One sentence may (and usually should) produce **MULTIPLE** spans.

Granularity rule: Prefer word-level or short phrase-level spans. If a shorter sub-phrase can carry the claim, use it instead of a longer one.

Recall rule: When uncertain whether something is a claim, **EXTRACT IT**. Over-extraction is strongly preferred to under-extraction.

VERIFICATION PRIORITY

1. Image content (**highest priority**)
 2. Instruction text
 3. Established factual or commonsense knowledge
- If there is a conflict:** Image content always prevails over text or prior knowledge.

CONFLICT CLASSIFICATION

Label **EACH** extracted span as exactly **ONE** of:

- **“Supported”** – The span is correct and consistent with the image, instruction, and factual knowledge.
- **“Vision-Prior Conflict”** – The span contradicts the image due to reliance on learned priors.
- **“Prior-Text Conflict”** – The span follows or accepts incorrect assumptions in the instruction that contradict factual knowledge.
- **“Vision-Text Conflict”** – The span reflects a mismatch between the image and the instruction.

QUALITY CONTROL (MANDATORY SELF-CHECK)

Before producing the final output, check:

1. Did you extract spans for: objects? attributes? quantities? spatial relations? actions?
2. Is the number of spans proportional to the length and detail of the completion?
 - A detailed paragraph should yield **MANY** spans.
 - Very few spans usually indicate missed conflicts.

OUTPUT FORMAT (STRICT)

Return **ONLY** a JSON array. Do **NOT** include explanations outside the JSON. The spans **MUST** be ordered by their first appearance in the completion.

```
[
  {
    "span": "exact minimal span from the completion",
    "label": "Supported | Vision-Prior Conflict | Prior-Text Conflict | Vision-Text Conflict"
  },
  ...
]
```

Box: Knowledge Conflict Annotation Prompt.

This prompt directs the judge model to decompose VLM responses into atomic spans and assign conflict labels (C_{VP}^e , C_{VT}^e , C_{PT}^e), providing granular

token-level supervision for conflict probe training.

B Implementation Details (Backbones and Decoding)

We evaluate three state-of-the-art reasoning-capable MLLMs spanning two architecture families and three distinct training recipes. Table 6 summarizes their key characteristics.

Table 6: Summary of evaluated MLLM backbones. The “Probe Layer” indicates the layer within the conflict-sensitive band used for probe training and intervention (identified in Section 4.3).

Attribute	R1-Onevision	Ocean-R1	Llama-3.2V-cot
Params	7B	7B	11B
Layers	28	28	40
Hidden Dim	3584	3584	4096
Probe Layer	20	20	39
Base Model	Qwen2.5-VL	Qwen2.5-VL	Llama-3.2-Vision

Ocean-R1 (Ming et al., 2025). Ocean-R1 is a vision-language reasoning model fine-tuned from Qwen2.5-VL-Instruct (Bai et al., 2025), optimized via a two-stage rule-based Reinforcement Learning (RL) framework. The first stage enhances reasoning capabilities, while the second refines visual perception. Empirical results on benchmarks such as MathVision and MathVista demonstrate its superior multimodal reasoning and generalization performance.

Llama-3.2V-11B-cot (Xu et al., 2025). Llama-3.2V-11B-cot is a reasoning-intensive model built on the Llama-3.2-Vision-11B architecture. Adopting the LLaVA-CoT methodology (Xu et al., 2025), it is fine-tuned on visual reasoning chains to perform step-by-step decomposition of complex queries. This Chain-of-Thought (CoT) mechanism enables precise handling of multi-hop reasoning and intricate scene understanding, dealing effectively with challenges where standard instruction-tuned models may falter.

R1-Onevision (Yang et al., 2025). R1-Onevision bridges visual perception and deep reasoning by fine-tuning Qwen2.5-VL (Bai et al., 2025) through a two-stage strategy combining Supervised Fine-Tuning (SFT) and RL. It introduces a cross-modal formalization pipeline that converts images into structured textual representations, complemented by a “role-playing” mechanism for iterative visual refinement. This

approach achieves state-of-the-art performance on benchmarks like MathVista and MathVerse, surpassing models such as GPT-4o.

Decoding Configuration. Unless otherwise specified, we use greedy decoding for all interventions to avoid confounding effects from sampling temperature. For methods requiring top- k candidates (e.g., probe-guided control; Appendix D.4.3), we set $k = 5$ uniformly across backbones.

C Probe Training and Architecture Details

This section provides the complete training configuration and architectural specifications for the conflict probes used in Section 4.

C.1 Probe Architectures

Canonical class-ID mapping. Table 7 fixes the correspondence between probe class IDs and conflict types. This mapping is used consistently throughout the paper, this appendix, and the released code (label converter, token labels, loss weighting, and metrics).

Table 7: Canonical mapping between probe class IDs and conflict types.

Class ID z	0	1	2	3
Meaning	Supported	$\mathcal{C}_{VP}$	$\mathcal{C}_{PT}$	$\mathcal{C}_{VT}$

We design two probe architectures to assess the linearity of conflict representations:

Linear Probe (f_{lin}). The linear probe consists of a single fully-connected layer that directly maps hidden states to conflict class logits:

$$f_{lin}(\mathbf{h}) = \mathbf{W}\mathbf{h} + \mathbf{b}, \quad (10)$$

Variable Definitions:

- $\mathbf{h} \in \mathbb{R}^d$: The input hidden state vector extracted from a specific layer of the MLLM at a given token position.
- $\mathbf{W} \in \mathbb{R}^{4 \times d}$: The learnable weight matrix that projects the d -dimensional hidden state to a 4-dimensional output (one dimension per class).
- $\mathbf{b} \in \mathbb{R}^4$: The learnable bias vector.
- d : The hidden state dimension, which varies by backbone architecture: $d = 3584$ for R1-Onevision-7B and Ocean-R1-7B, and $d = 4096$ for Llama-3.2V-11B-cot.

- Output dimension 4 corresponds to the four classes in the canonical mapping of Table 7: {No-Conflict, $\mathcal{C}_{VP}^e$, $\mathcal{C}_{PT}^e$, $\mathcal{C}_{VT}^e$ }.

Rationale: The linear probe tests whether conflict states are *linearly separable* in the hidden state space. High accuracy with a linear probe indicates that conflicts are explicitly encoded as directions in the representation space, rather than being entangled in complex nonlinear manifolds. This is crucial for our intervention methods, which rely on linear steering vectors.

MLP Probe (f_{mlp}). The MLP probe uses a three-layer architecture with progressive dimensionality reduction, designed to capture potential nonlinear decision boundaries:

$$\mathbf{u}_1 = \text{ReLU}(\mathbf{W}_1\mathbf{h} + \mathbf{b}_1), \quad \mathbf{W}_1 \in \mathbb{R}^{1024 \times d}, \quad (11)$$

$$\mathbf{u}_2 = \text{ReLU}(\mathbf{W}_2\mathbf{u}_1 + \mathbf{b}_2), \quad \mathbf{W}_2 \in \mathbb{R}^{512 \times 1024}, \quad (12)$$

$$\mathbf{u}_3 = \text{ReLU}(\mathbf{W}_3\mathbf{u}_2 + \mathbf{b}_3), \quad \mathbf{W}_3 \in \mathbb{R}^{256 \times 512}, \quad (13)$$

$$f_{mlp}(\mathbf{h}) = \mathbf{W}_4\mathbf{u}_3 + \mathbf{b}_4, \quad \mathbf{W}_4 \in \mathbb{R}^{4 \times 256}. \quad (14)$$

Variable Definitions:

- $\mathbf{h} \in \mathbb{R}^d$: The input hidden state vector (same as linear probe).
- $\mathbf{u}_1 \in \mathbb{R}^{1024}$: First hidden layer output after ReLU activation.
- $\mathbf{u}_2 \in \mathbb{R}^{512}$: Second hidden layer output after ReLU activation.
- $\mathbf{u}_3 \in \mathbb{R}^{256}$: Third hidden layer output after ReLU activation.
- $\mathbf{W}_i, \mathbf{b}_i$: Learnable weight matrices and bias vectors for layer i .
- $\text{ReLU}(\cdot) = \max(0, \cdot)$: Rectified Linear Unit activation function, introducing nonlinearity.

Architecture Design: The progressive dimensionality reduction ($d \rightarrow 1024 \rightarrow 512 \rightarrow 256 \rightarrow 4$) acts as an information bottleneck, forcing the probe to extract the most discriminative features for conflict classification. We apply dropout with probability $p = 0.1$ after each ReLU activation during training to prevent overfitting.

Comparison Purpose: Comparing the MLP probe’s performance against the linear probe reveals whether conflict information is linearly encoded. In our experiments, the marginal improvement of MLP over linear probes (<2% AUC gain)

confirms that conflicts are predominantly linearly separable.

C.2 Training Configuration

Data split. We use an 80/20 train/test split for each backbone, as summarized in Table 5. Hidden states are extracted from pre-computed reasoning trajectories and cached to disk for efficient training.

Optimization. We train all probes using AdamW with the following hyperparameters:

- Learning rate: 1×10^{-3} (linear probe), 5×10^{-4} (MLP probe)
- Weight decay: 1×10^{-4}
- Batch size: 256 tokens
- Training epochs: 10 (with early stopping based on validation loss, patience = 3)
- Learning rate scheduler: Cosine annealing with warm-up (5% of total steps)

Class weighting. To address the extreme class imbalance in long-CoT reasoning (conflict tokens are sparse, typically $<5\%$ of all tokens), we employ inverse-frequency weighting in the cross-entropy loss:

$$w_z = \frac{N}{\sum_t \mathbb{I}[z_t = z] + \epsilon}, \quad (15)$$

Variable Definitions:

- w_z : The weight assigned to class z in the weighted cross-entropy loss. Higher weights penalize misclassification of rare classes more heavily.
- N : The total number of tokens in the training set.
- $z_t \in \{0, 1, 2, 3\}$: The ground-truth class label for token t , following Table 7: 0 = No-Conflict, 1 = $\mathcal{C}_{VP}^e$, 2 = $\mathcal{C}_{PT}^e$, 3 = $\mathcal{C}_{VT}^e$.
- $\mathbb{I}[\cdot]$: The indicator function, which equals 1 if the condition is true and 0 otherwise.
- $\sum_t \mathbb{I}[z_t = z]$: The count of tokens belonging to class z in the training set.
- $\epsilon = 100$: A smoothing constant to prevent division by zero and to moderate extreme weights for very rare classes.

Intuition: Classes with fewer samples receive higher weights, ensuring the probe does not trivially predict the majority class (No-Conflict). In practice, this results in approximate weight ratios of 1 : 8 : 12 : 10 for $z \in \{0, 1, 2, 3\}$ (No-Conflict : $\mathcal{C}_{VP}^e$: $\mathcal{C}_{PT}^e$: $\mathcal{C}_{VT}^e$), though exact values vary by backbone due to different conflict distributions in their reasoning trajectories.

Layer selection for training. For the main experiments in Section 4.4, we train probes within the identified *conflict-sensitive layer band* (Section 4.3):

- R1-Onevision-7B: Layer 20
- Ocean-R1-7B: Layer 20
- Llama-3.2V-11B-cot: Layer 39

For the layer-scanning experiments (Section 4.3), we train separate probes for each layer $l \in \{1, 2, \dots, L\}$ using identical hyperparameters.

C.3 Attention Activation Analysis

Activation threshold (γ). For the head activation ratio analysis in Section 4.3, we set the activation threshold γ based on the median ℓ_2 -norm of attention head outputs across the training set:

$$\gamma = \text{median}_{(x,t,l,a)} \left\| \mathbf{o}_t^{(l,a)} \right\|_2. \quad (16)$$

Variable Definitions:

- γ : The activation threshold used to determine whether an attention head is “active” (contributing significantly to the output).
- $\mathbf{o}_t^{(l,a)} \in \mathbb{R}^{d_{\text{head}}}$: The output vector of attention head a at layer l for token position t .
- $\|\cdot\|_2$: The ℓ_2 (Euclidean) norm, measuring the magnitude of the head output vector.
- $\text{median}_{(x,t,l,a)}$: The median taken over all samples x , token positions t , layers l , and attention heads a in the training set.
- d_{head} : The dimension of each attention head output (typically d/n_{heads}).

Usage: An attention head is considered “activated” at position t if $\left\| \mathbf{o}_t^{(l,a)} \right\|_2 > \gamma$. We then compute the *Head Activation Ratio* (HAR) as the fraction of heads exceeding this threshold, which serves as a proxy for the model’s internal “engagement” during conflict processing.

Calibration: Using the median ensures γ is data-driven and robust to outliers. This yields $\gamma \approx 0.15$ for R1-Onevision and Ocean-R1, and $\gamma \approx 0.12$ for Llama-3.2V. We verified that our conclusions are robust to $\pm 20\%$ perturbations in γ .

C.4 Visualization Parameters

t-SNE configuration. For the sample-level separability visualizations in Figure 3, we use the following t-SNE parameters:

- Perplexity: 30
- Learning rate: 200
- Number of iterations: 1000
- Initialization: PCA

- Random seed: 42 (for reproducibility)

Sample-level representations are obtained by mean-pooling hidden states across all tokens where the probe predicts an active conflict ($\hat{z}_t \neq 0$).

C.5 Aggregation Strategies

We evaluate probe performance at three granularities reported in Table 2:

All Token. Evaluation on all valid tokens (excluding padding), measuring fine-grained per-token detection.

Span Only. Evaluation restricted to tokens within annotated spans, measuring intra-span detection accuracy.

Span-Max. For each annotated span, we aggregate token-level predictions by taking the maximum predicted conflict probability:

$$P_{\text{span}} = \max_{t \in \text{span}} P_{\phi}(z_{\text{span}} | \mathbf{h}_t^{(l)}), \quad (17)$$

Variable Definitions:

- $P_{\text{span}} \in [0, 1]$: The aggregated span-level probability that this span exhibits the ground-truth conflict type.
- $t \in \text{span}$: Token indices belonging to the annotated span (e.g., if a span covers tokens 15–20, then $t \in \{15, 16, 17, 18, 19, 20\}$).
- $P_{\phi}(z_{\text{span}} | \mathbf{h}_t^{(l)})$: The probe’s predicted probability for the ground-truth conflict class z_{span} , given the hidden state $\mathbf{h}_t^{(l)}$ at token t and layer l . This is obtained by applying softmax to the probe output:

$$P_{\phi}(z | \mathbf{h}) = \frac{\exp(f_{\phi}(\mathbf{h})_z)}{\sum_{z'} \exp(f_{\phi}(\mathbf{h})_{z'})} \quad (18)$$

- $z_{\text{span}} \in \{1, 2, 3\}$: The ground-truth conflict label for this span ($1 = C_{\text{VP}}^e$, $2 = C_{\text{PT}}^e$, $3 = C_{\text{VT}}^e$; Table 7).
- $\mathbf{h}_t^{(l)} \in \mathbb{R}^d$: The hidden state at layer l and token position t .

Intuition: Taking the maximum across tokens captures the “peak signal” within a span. This is motivated by the observation that conflict signals often concentrate at specific “pivot” tokens (e.g., the critical word where the model commits to an incorrect claim), rather than being uniformly distributed across all span tokens.

Evaluation: A span is classified as conflicted if $P_{\text{span}} > 0.5$, yielding a span-level binary judgment.

Span-Max consistently outperforms other granularities (see Table 2), confirming that conflict information is spatially concentrated. All main results in Section 4 use Span-Max aggregation.

C.6 Layer-Selection Sensitivity

Section 4.3 selects a single probing layer per backbone from the conflict-sensitive band (Layer 20 for the 7B models, Layer 39 for Llama-3.2V). Because this choice could in principle be doing the work, we audit every layer of every backbone and compare the layer we use against the best-performing layer, averaging one-vs-rest AUC over C_{VP} , C_{PT} , and C_{VT} under the same protocol.

Table 8: Layer-selection sensitivity: mean one-vs-rest AUC (%) at the layer used in the paper versus the best layer found by an exhaustive scan.

Model	Paper layer	Best layer	Gap
Ocean-R1	97.76	97.91	0.15
R1-Onevision	96.70	97.40	0.70
Llama-3.2V	96.34	97.35	1.01

As Table 8 shows, the penalty for our fixed choice is at most 1.01 AUC points, and strong performance extends over a broad mid-to-late region rather than a single privileged layer (Figure 4). We therefore describe the result as a robust, backbone-dependent *band* of depth-sensitive readability, and none of our conclusions depend on selecting the single optimal layer.

C.7 Cross-Model Consistency of Conflict Encoding

A natural concern is whether the observed patterns are artifacts of the data construction process (e.g., biases introduced by closed-source annotation models) rather than properties of the models themselves. To address this, we examine the consistency of our findings across three trained backbones.

Backbone coverage. Our evaluation spans *two* architecture families and three training recipes: (1) **Qwen2.5-VL-based models** (R1-Onevision-7B, Ocean-R1-7B): 7B parameters, 28 layers, sharing a base model but trained with different SFT/RL strategies; (2) **Llama-3.2-Vision-based model** (Llama-3.2V-11B-cot): 11B parameters, 40 layers, with a different vision encoder and pretraining corpus. We therefore describe our evidence as replication across three trained models

from two architecture families, since variation in training recipe is training-level rather than architectural evidence, and a genuinely third architecture family remains an extension. Across this coverage, we observe consistent patterns.

Consistent findings across backbones. Table 9 summarizes key metrics across the three backbones. Several observations support the robustness of our conclusions:

- **Linear separability:** All three models exhibit high linear probe AUC (93–99%), indicating that conflict states are explicitly encoded rather than implicitly entangled, regardless of architecture.
- **Depth localization:** Conflict signals consistently peak in mid-to-late layers (L15–22 for 7B models, L30–39 for 11B), following a similar relative depth pattern (~ 70 –80% of total layers).
- **Conflict-type ranking:** $\mathcal{C}_{PT}$ is consistently the most separable (AUC > 98%), while $\mathcal{C}_{VP}$ is the most challenging across all backbones, suggesting this ranking reflects inherent task difficulty rather than annotation bias.
- **Intervention response:** All models respond to probe-guided control with 60–80% conflict rate reduction, and all exhibit directional asymmetry favoring forward steering.

Table 9: Cross-model consistency of conflict encoding patterns. Despite architectural differences, all three backbones exhibit similar separability, depth localization, and intervention response.

Model	Linear AUC	Peak Layer	$\mathcal{C}_{PT}$ AUC	$\mathcal{C}_{VP}$ AUC	CR Red.
R1-Onevision	96.2%	L20 (71%)	98.6%	93.4%	80%
Ocean-R1	95.8%	L20 (71%)	98.6%	93.2%	76%
Llama-3.2V	97.1%	L39 (78%)	99.2%	95.9%	67%

This cross-model consistency indicates that the observed conflict-encoding patterns are consistent across backbones and training recipes rather than artifacts of our specific data construction or annotation pipeline. It does not by itself establish that the same underlying circuit is at work in each model.

C.8 Mechanistic Heterogeneity of Conflict Types

While we adopt a unified probing framework to study all three conflict types, we emphasize that this does *not* imply their underlying mechanisms are identical. Rather, the unified framework enables systematic comparison and reveals both commonalities and differences.

Distinct cognitive demands. The three conflict types impose fundamentally different cognitive demands on the model:

- **$\mathcal{C}_{VP}$ (Vision-Prior Conflict):** Requires the model to reconcile counter-intuitive visual input with internalized world knowledge. This is a *perceptual calibration* setting, in which parametric memory competes with visual features from the vision encoder.
- **$\mathcal{C}_{PT}$ (Prior-Text Conflict):** Requires the model to resist misleading textual claims that contradict its factual knowledge. This is a *factual grounding* task, primarily involving competition between in-context information and parametric knowledge retrieval.
- **$\mathcal{C}_{VT}$ (Vision-Text Conflict):** Requires the model to arbitrate between contextual textual information and visual evidence. This resembles an *instruction-following under distraction* task, where the model must resist misleading context while remaining grounded in visual input. This setting is dominated by competition in cross-modal attention.

Evidence for type-specific structure. Our empirical results reveal several signatures of these differences:

- **Separability hierarchy:** $\mathcal{C}_{PT}$ achieves near-perfect probe accuracy (99.4% conditioned recall), while $\mathcal{C}_{VP}$ is more challenging (80.7%). This aligns with the intuition that text-prior conflicts produce clearer decision boundaries, whereas vision-prior conflicts involve more subtle perceptual-semantic interactions.
- **Confusion patterns:** As shown in Figure 3, the primary confusion occurs between $\mathcal{C}_{VP}$ and $\mathcal{C}_{VT}$ (both involving visual processing failures), while $\mathcal{C}_{PT}$ forms a distinctly separated cluster. This suggests that vision-related conflicts share partially overlapping representations.
- **Layer-wise encoding:** Figure 4 shows that different conflict types peak at slightly different layers, with $\mathcal{C}_{PT}$ typically achieving peak separability 1–2 layers earlier than visual conflicts, consistent with the hypothesis that text-prior arbitration occurs upstream of visual integration.

Value of a unified framework. Despite these differences, a unified probing approach offers several advantages:

- **Comparable diagnosis:** Using identical probe architectures enables fair comparison of encoding strength and intervention efficacy across con-

flict types.

- **Shared intervention substrate:** All three conflict types respond to probe-guided control (Table 4), suggesting that while the underlying mechanisms differ, they converge to a shared representational substrate that can be leveraged for unified control.
- **Practical utility:** A single probe can detect all conflict types in real-time, enabling a unified diagnostic tool for long-CoT reasoning failures regardless of the specific conflict source.

The clear separability observed in Figure 3 should therefore be interpreted not as evidence against a unified framework, but rather as validation that our probe successfully captures the distinct signatures of each conflict type while enabling systematic cross-type comparison.

C.9 Stricter Online Detection Settings

Although Span-Max aggregation in Section 4.4 yields the strongest separability, it is not strictly equivalent to fully online detection: it pools predictions across all tokens within an annotated span. To assess how much conflict signal is recoverable under *stricter prefix-only protocols*, we evaluate two additional settings.

First-token setting. For each conflict span, only the prediction at the *first* token of the span is used. This is the most adversarial online setting: the probe must commit to a conflict label using only the earliest, locally weakest evidence in the span.

Fixed-window setting ($w = 8$). A prefix-only sliding window of length $w=8$ tokens: when the running window first enters a span, we use the probe prediction from that window. This emulates a realistic streaming detector that triggers on the leading edge of a conflict rather than the full span.

Table 10 compares these protocols to the All Token, Span Only, and Span-Max settings already reported in Table 2 (averaged across conflict types and backbones).

Table 10: Comparison of token-level / online aggregation settings (avg. across C_{VP} , C_{PT} , C_{VT} and three backbones).

Setting	Avg. AUC (%)	Avg. Recall@0.1 (%)
first_token	89.8	71.8
fixed-window	95.1	85.1
Span-Max	97.8	94.7

Even the most adversarial first_token (on-

line, strict) setting achieves 89.8% AUC, and the strictly online fixed-window ($w=8$, online) setting reaches 95.1% AUC. The remaining gap to Span-Max (reference) is consistent with our observation in Section 4.2 that conflict spans are *semantically sparse*: many internal tokens carry only weak local evidence, so Span-Max aggregation acts as a denoiser rather than as a fundamentally different mechanism. We therefore interpret “streaming diagnosis” as the property that *prefix-only signals already contain substantial conflict information*, while Span-Max gives the cleanest measurement when offline labels are available.

C.10 Surface-Signal Control Baselines

A strong probe alone does not rule out the possibility that detection exploits superficial cues correlated with conflict (e.g., uncertain tokens, late positions). To address this, we compare the probe to three control baselines under the *same* train/test split, conflict labels, and aggregation protocols as Table 2:

- **Position-only:** features are token-position embeddings only.
- **Token-entropy:** per-token next-token entropy of the base model.
- **Perplexity:** running per-token perplexity of the base model.

Table 11: Span-Max AUC (%) of surface-signal control baselines under the same protocol as Table 2.

Baseline	w/o Conf.	$C_{VP}^o=1$	$C_{PT}^o=1$	$C_{VT}^o=1$
Position	45.3	46.4	67.1	67.5
Token Entropy	53.1	88.9	69.2	59.0
Perplexity	52.4	47.9	63.4	68.5
Probe (ours)	93.9	93.4	98.4	97.2

Table 12: All-Token AUC (%) of surface-signal control baselines (same protocol).

Baseline	w/o Conf.	$C_{VP}^o=1$	$C_{PT}^o=1$	$C_{VT}^o=1$
Position	67.3	63.9	75.5	66.6
Token Entropy	51.8	62.7	53.2	53.8
Perplexity	51.2	40.2	52.2	45.8
Probe (ours)	88.7	90.5	96.9	94.5

The patterns are consistent: *position* is weak and inconsistent; *perplexity* is uniformly weak; *entropy* captures partial uncertainty for C_{VP}^o but fails to separate other conflict types. None of these surface cues comes close to the probe’s balanced performance across all conflict types. While individual

surface signals may correlate with conflict in isolated cases, they cannot explain the overall pattern, indicating that the probe is detecting structured conflict-related representations rather than exploiting shallow shortcuts.

C.11 Cross-Domain (OOD) Robustness

To assess applicability beyond our constructed benchmark, we build an OOD stress test sampling 100 challenging long-CoT cases (20 per dataset) from five heterogeneous benchmarks: MMMU (Yue et al., 2025), MATHVISTA, CHARTQA, HALLUSIONBENCH (Guan et al., 2024), and MEDVQA. These cover distinct task formats (knowledge-intensive QA, mathematical reasoning, chart understanding, hallucination probing, and medical VQA) and were *not* seen during probe training.

Probe performance under distribution shift. Using Span-Max aggregation, the same probe (without re-training) yields:

Table 13: OOD evaluation aggregated across five external benchmarks.

Metric	Value (%)
Accuracy	66.03
Macro-F1	56.69
OvR AUC (4-class)	86.41
Binary Accuracy	69.17
Binary AUC	87.15
Recall@FPR=0.1	79.80

The Binary AUC of 87.15% indicates the probe retains strong conflict-vs-no-conflict separability even under substantial distribution shift, supporting the claim that conflict signals are not artifacts of our construction pipeline.

OOD confusion structure. Table 14 reports the OOD confusion matrix. C_{PT}^o is almost perfectly identified, while most errors fall into the w/o Conflict class for visually-grounded types (C_{VP}^o , C_{VT}^o), consistent with our in-distribution observation that visual conflicts are semantically sparser.

Table 14: OOD confusion matrix (Gold rows, Predicted columns).

Gold \ Pred	w/o Conf.	$C_{VP}^o=1$	$C_{PT}^o=1$	$C_{VT}^o=1$
w/o Conflict	327	3	10	1
$C_{VP}^o=1$	158	47	15	3
$C_{PT}^o=1$	31	1	99	0
$C_{VT}^o=1$	23	1	3	11

These results suggest the probe captures *structured conflict signals* that transfer across task formats, rather than exploiting templates specific to our construction.

C.12 Annotation Pipeline Validity

To further address concerns that probe performance might reflect annotation artifacts rather than genuine conflict signals, we conduct a focused human evaluation on 100 randomly sampled trajectories (covering all three conflict types and three backbones). Two independent annotators, blind to model identity and pipeline labels, were asked to (i) identify all conflict locations in the trajectory and (ii) assign a conflict type to each.

- **Conflict localization:** 93% of spans flagged by the pipeline were confirmed as actual conflicts by both annotators, with no missed conflicts in the verified subset.
- **Conflict typing:** 94.6% inter-rater agreement with the automated typing.

This complements the agreement statistics for the LLM-as-judge in Appendix D.2 ($\kappa=0.87$ on $\sim 1,500$ spans) and supports that the supervision signal used to train the probe is consistent with human judgment.

D Directional Intervention and Evaluation

This appendix provides full definitions, hyperparameters, and prompts for Section 5.

D.1 Notation and control settings

Inputs, outputs, and trajectories. We use the same notation as Sec. 3. An input is $x = (X_V, X_T)$, where X_V is an image and X_T is the textual input (context/instruction). The model M_θ generates an output sequence $y = (y_1, \dots, y_T)$. We denote the hidden state at layer l and step t as $\mathbf{h}_t^{(l)}$.

Objective vs. effective conflict. $C_{i,j}^o(x) = 1$ indicates that sources i and j disagree factually on input x . $C_{i,j}^e(t | x) = 1$ indicates that at generation step t the model exhibits conflict-induced behavior consistent with an active conflict between the same sources.

Forward vs. Reverse directions. Table 15 defines **Forward** vs. **Reverse** control for each conflict type. In words: **Forward** follows the dataset-level reliability prior ($VP/VT \rightarrow \mathcal{K}_{\text{vision}}$,

$PT \rightarrow \mathcal{K}_{\text{prior}}$); **Reverse** follows the opposite source ($VP \rightarrow \mathcal{K}_{\text{prior}}$, $VT/PT \rightarrow \mathcal{K}_{\text{text}}$).

Table 15: Forward and Reverse direction mappings for each conflict type. **Forward** aligns with the more reliable source based on our data construction. **Reverse** enforces the opposite choice.

Type	Conflicting Sources	Forward ($\uparrow$)	Reverse ($\downarrow$)
$\mathcal{C}_{VP}$	Vision vs. Prior	$\mathcal{K}_{\text{vision}}$	$\mathcal{K}_{\text{prior}}$
$\mathcal{C}_{VT}$	Vision vs. Text	$\mathcal{K}_{\text{vision}}$	$\mathcal{K}_{\text{text}}$
$\mathcal{C}_{PT}$	Prior vs. Text	$\mathcal{K}_{\text{prior}}$	$\mathcal{K}_{\text{text}}$

Why no VCD for $\mathcal{C}_{PT}$ subset in Figure 5 and Table 4? While VCD is a useful decoding-time knob for visual conflicts, it is not naturally defined for $\mathcal{C}_{PT}$ (Prior-Text Conflict) in our setup, as VCD specifically targets vision-related conflicts by contrasting visual and non-visual distributions. To keep the mitigation comparison complete across the three conflict subsets, we therefore report only steering and probe-guided results for the $\mathcal{C}_{PT}$ subset in this analysis.

D.2 LLM-as-a-Judge Pipeline

We implement an automated semantic evaluation pipeline used to produce the results in Figure 5. The pipeline decomposes model responses into verifiable atomic claims and checks each claim against a target truth anchor.

Truth anchors. The truth anchor depends on the intended target source $\mathcal{K}_s$:

- If $s = \text{vision}$, the anchor is the image X_V (we use an MLLM judge that can condition on the image).
- If $s = \text{text}$, the anchor is the textual input X_T (the judge is required to cite a supporting span from X_T).
- If $s = \text{prior}$, the anchor is world knowledge. In our setup, the judge uses its internal knowledge and is allowed to output unknown when unsure. (Optional: retrieval augmentation can be plugged in, but is not required for our analysis.)

Atomic claim extraction. Given the model response y , the judge extracts a list of atomic claims $\mathcal{Q}(y) = \{a_m\}_{m=1}^M$ such that each a_m is: (i) a single factual assertion, (ii) minimally checkable, (iii) not a subjective or stylistic statement.

Verification labels. For each claim a_m , the judge outputs a label

$$r_m \in \{\text{supported}, \text{contradicted}, \text{unknown}\}. \quad (19)$$

We explicitly keep unknown to prevent over-penalizing claims that are not decidable from the anchor (e.g., the image is blurry, the text is under-specified, or the world-knowledge check is uncertain).

Aggregation metrics. Given the set of verification labels $\{r_m\}_{m=1}^M$, we compute three semantic evaluation metrics:

$$\text{ASR} = \frac{1}{M} \sum_{m=1}^M \mathbb{I}[r_m = \text{supported}], \quad (20)$$

$$\text{ARR} = \frac{1}{M} \sum_{m=1}^M \mathbb{I}[r_m = \text{contradicted}]. \quad (21)$$

Variable Definitions:

- **ASR (Anchor Support Rate):** The proportion of atomic claims that are verified as consistent with the target truth anchor. Higher ASR indicates better alignment with the intended knowledge source.
- **ARR (Anchor Refutation Rate):** The proportion of atomic claims that directly contradict the target truth anchor. Lower ARR indicates fewer factual errors.
- M : The total number of extracted atomic claims from the model response.
- $r_m \in \{\text{supported}, \text{contradicted}, \text{unknown}\}$: The verification label for claim a_m .
- $\mathbb{I}[\cdot]$: The indicator function.

To quantify “obvious” semantic errors (where the judge is highly confident), the judge additionally outputs a confidence score $c_m \in [0, 1]$ for each verdict. We then define the Obvious Error Rate:

$$\text{OER} = \frac{1}{M} \sum_{m=1}^M \mathbb{I}[r_m = \text{contradicted} \wedge c_m \geq c_{\text{thresh}}], \quad (22)$$

Additional Variables:

- **OER (Obvious Error Rate):** The proportion of claims that are both contradicted *and* judged with high confidence. This filters out ambiguous cases and focuses on clear-cut errors.
- $c_m \in [0, 1]$: The judge’s confidence score for claim a_m . Higher values indicate greater certainty.

Algorithm 1: Extract–Verify–Score (LLM-as-a-Judge)

Input: Input $x = (X_V, X_T)$, target source $\mathcal{K}_s$, model response y
Output: ASR, ARR, OER
 $\mathcal{Q}(y) \leftarrow \text{EXTRACTATOMICCLAIMS}(y)$;
foreach $a \in \mathcal{Q}(y)$ **do**
 $(v, c) \leftarrow$
 $\text{VERIFYAGAINSTANCHOR}(a, x, \mathcal{K}_s)$;
 store (v, c) ;
Compute ASR/ARR/OER by aggregating $\{r_m\}$;
Compute OER from $\{(v, c)\}$ using Equation (22);

- c_{thresh} : A fixed confidence threshold (we use $c_{\text{thresh}} = 0.8$ in all experiments).
- $\wedge$: Logical AND operator. Both conditions must be satisfied.

Interpretation: A successful intervention should *increase* ASR (more claims align with the target source) and *decrease* ARR and OER (fewer contradictions). The gap between Forward and Reverse directions quantifies the model’s controllability.

Self-consistency (optional). To reduce judge variance, we optionally run the judge multiple times with the same prompt and aggregate verdicts by majority vote (ties resolved as unknown). All our main conclusions are robust to this choice.

Prompt templates. We keep the judge prompts fixed across methods and backbones. For the text-anchor case, the judge must return a quoted span from X_T when predicting supported. For the vision-anchor case, the judge must describe the visual evidence that supports or refutes the claim. We include the full prompts in our release.

D.3 Token-level Probe Metrics

This section provides the complete definitions of the token-level metrics reported in Section 5.3 and Table 4. Table 16 provides a quick reference.

Probe output. At each token step t , the probe outputs a distribution $P_t \in \mathbb{R}^4$ over labels $z \in \{0, 1, 2, 3\}$ (Section 4). We denote $P_t(0)$ as the probability of no-conflict. Let $\hat{z}_t = \arg \max_z P_t(z)$ be the hard prediction.

Micro-aggregation protocol. We only evaluate assistant-generated tokens (excluding system/user

Table 16: Summary of token-level probe metrics used for conflict mitigation evaluation.

Metric	Dir.	Interpretation
SS (Support)	↑	Avg. prob. of no-conflict state
CR (Conflict Rate)	↓	Fraction of conflict-predicted tokens
CAC (Conf.-Adj.)	↓	Penalizes confident conflict preds.
CCI (Conf. Index)	↓	Avg. confidence in conflict state

prompts). We aggregate across all valid tokens in the entire evaluation split (equivalent to concatenating all outputs into one long token stream), which stabilizes estimates under long-CoT length variation.

Support Score (SS).

$$\text{SS} = \frac{1}{N} \sum_{t=1}^N P_t(0). \quad (23)$$

Variable Definitions:

- $\text{SS} \in [0, 1]$: Support Score, the average probability that tokens are in a no-conflict (supported) state.
- N : Total number of evaluated tokens (assistant-generated tokens only).
- $P_t(0)$: The probe’s predicted probability that token t belongs to class 0 (no-conflict).

Interpretation: SS is high when most tokens are predicted to be supported (no effective conflict). A successful Forward intervention should *increase* SS.

Conflict Rate (CR).

$$\text{CR} = \frac{1}{N} \sum_{t=1}^N \mathbb{I}[\hat{z}_t \neq 0]. \quad (24)$$

Variable Definitions:

- $\text{CR} \in [0, 1]$: Conflict Rate, the fraction of tokens with an active predicted conflict.
- $\hat{z}_t = \arg \max_z P_t(z)$: The hard (argmax) prediction for token t .
- $\mathbb{I}[\hat{z}_t \neq 0]$: Indicator that equals 1 if the predicted class is any conflict type (1, 2, or 3), and 0 if no-conflict (class 0).

Interpretation: CR measures the fraction of tokens at which the probe predicts an active effective conflict. A successful Forward intervention should *decrease* CR.

Confidence-adjusted conflict (CAC). First, we define two intermediate quantities:

Conflict mass: $q_t = 1 - P_t(0) \in [0, 1]$, representing the total probability mass assigned to any conflict class.

Normalized entropy:

$$H_n(P_t) = \frac{-\sum_{z=0}^3 P_t(z) \log P_t(z)}{\log 4} \in [0, 1], \quad (25)$$

where the denominator $\log 4$ normalizes to $[0, 1]$ (maximum entropy over 4 classes).

Certainty: $u_t = 1 - H_n(P_t) \in [0, 1]$, representing how confident/certain the probe is (high when the distribution is peaked).

The CAC metric is then defined as:

$$\text{CAC} = \frac{1}{N} \sum_{t=1}^N q_t \cdot u_t. \quad (26)$$

Interpretation: CAC penalizes tokens that are both conflict-heavy (large q_t) and confident (large u_t), corresponding to “confident conflict.” This metric is more nuanced than CR because it down-weights uncertain conflict predictions.

Conflict confidence index (CCI). Let $\mathcal{I}_{\text{conf}} = \{t \mid \hat{z}_t \neq 0\}$ be the set of token indices where the probe predicts a conflict.

$$\text{CCI} = \frac{1}{|\mathcal{I}_{\text{conf}}|} \sum_{t \in \mathcal{I}_{\text{conf}}} P_t(\hat{z}_t). \quad (27)$$

Variable Definitions:

- $\text{CCI} \in [0, 1]$: Conflict Confidence Index, the average confidence of conflict predictions.
- $|\mathcal{I}_{\text{conf}}|$: The number of tokens where a conflict is predicted.
- $P_t(\hat{z}_t)$: The probability assigned to the predicted conflict class at token t .

Interpretation: CCI measures how “stubborn” the model is when it enters a conflict state: higher means the model assigns high confidence to a conflict class whenever it is in conflict. A successful intervention should *decrease* CCI, indicating the model is less committed to conflict-inducing behaviors.

Relationship between semantic and token-level metrics. We employ two complementary evaluation paradigms—semantic-level (ASR/ARR/OER) and token-level (SS/CR/CAC/CCI)—that capture different aspects of intervention effectiveness:

- **Semantic metrics** (ASR, ARR, OER) evaluate the *end result*: whether the final generated claims align with the target knowledge source. These metrics are directly interpretable but require expensive LLM-based judging and only assess the output, not the generation process.
- **Token-level metrics** (SS, CR, CAC, CCI) evaluate the *generation dynamics*: whether the model’s internal states exhibit conflict signals during generation. These metrics are computationally cheap (requiring only probe inference) and provide fine-grained insight into *when* and *where* conflicts occur.

D.4 Intervention implementations

We now detail the three inference-time interventions used in Section 5. All interventions are applied at inference time without updating θ .

D.4.1 VCD (Vision-Contrastive Decoding)

We adopt VCD (Leng et al., 2024) as a decoding-time logit intervention for conflicts involving vision. At each step t , VCD constructs two logit vectors and applies contrastive correction:

$$\tilde{\ell}_t = \ell_t^g + \beta (\ell_t^g - \ell_t^u), \quad (28)$$

Variable Definitions:

- $\ell_t^g \in \mathbb{R}^{|V|}$: The *grounded* logits vector computed on the full multimodal input (X_V, X_T) , where $|V|$ is the vocabulary size. This represents the model’s “normal” next-token distribution with full visual access.
- $\ell_t^u \in \mathbb{R}^{|V|}$: The *ungrounded* logits vector computed on a contrastive input where the visual signal is weakened (e.g., by masking the image, replacing it with a blank/noise image, or using only a textual summary). This represents the model’s prediction when it cannot rely on visual evidence.
- $\tilde{\ell}_t \in \mathbb{R}^{|V|}$: The corrected logits used for sampling the next token.
- $\beta > 0$: The correction strength hyperparameter. Larger β applies stronger visual grounding correction.

Intuition: The difference $(\ell_t^g - \ell_t^u)$ isolates the contribution of visual information to token predictions. Adding this difference (scaled by β) amplifies tokens that the model would predict *only when it has visual access*, thereby improving visual grounding. Conversely, tokens that are equally likely with or without vision (driven purely by prior/text) are not boosted.

Applicability. Because ℓ_t^u is defined by manipulating the visual channel, VCD is naturally aligned with conflicts where vision is one of the two sources ($\mathcal{C}_{VP}^e, \mathcal{C}_{VT}^e$). We therefore do not apply it to $\mathcal{C}_{PT}^e$ (Prior-Text Conflict), where no visual manipulation is meaningful.

D.4.2 Representation steering

Steering exploits the approximately linear structure of conflict representations discovered in Section 4.4. The key insight is that conflict states correspond to specific *directions* in the hidden state space, which can be directly manipulated.

Steering vector construction. For a chosen layer l (within the conflict-sensitive layer band of Section 4.3), we construct a direction vector by differencing mean hidden states:

$$\mathbf{v}^{(l)} = \mathbb{E}[\mathbf{h}^{(l)} \mid \text{target}] - \mathbb{E}[\mathbf{h}^{(l)} \mid \text{reference}]. \quad (29)$$

Variable Definitions:

- $\mathbf{v}^{(l)} \in \mathbb{R}^d$: The steering vector at layer l , representing the direction from reference behavior to target behavior in the hidden state space.
- $\mathbb{E}[\mathbf{h}^{(l)} \mid \text{target}]$: The mean hidden state at layer l across all tokens exhibiting the target behavior (e.g., vision-following, no-conflict).
- $\mathbb{E}[\mathbf{h}^{(l)} \mid \text{reference}]$: The mean hidden state at layer l across all tokens exhibiting the reference behavior (e.g., prior-following, in-conflict).

Semantic Interpretation of (target, reference):

- **Conflict mitigation:** target = no-conflict tokens, reference = conflict tokens. The vector $\mathbf{v}^{(l)}$ points “away from conflict.”
- **Forward steering:** target = tokens following the reliable source, reference = tokens following the unreliable source. The vector points toward factually correct behavior.
- **Source preference:** target = vision-following tokens, reference = text/prior-following tokens. The vector encodes visual grounding.

In all cases, we ℓ_2 -normalize $\mathbf{v}^{(l)}$ to unit length before use, ensuring consistent intervention magnitude.

Injection rule. During generation, we modify the hidden state at layer l by adding the steering vector:

$$\tilde{\mathbf{h}}_t^{(l)} = \mathbf{h}_t^{(l)} + \lambda \mathbf{v}^{(l)}. \quad (30)$$

Variable Definitions:

- $\tilde{\mathbf{h}}_t^{(l)}$: The modified hidden state at layer l , token t , which is passed to subsequent layers.
- $\mathbf{h}_t^{(l)}$: The original (unmodified) hidden state.
- $\lambda \in \mathbb{R}$: The steering strength. Positive λ moves the representation toward target. Negative λ moves toward reference. Typical values: $\lambda \in [0.3, 1.0]$.

Application Strategy: We apply this intervention either (a) at all decoding steps (unconditional steering), or (b) only when the probe predicts elevated conflict mass $q_t > \delta$ (conditional steering), where δ is a predefined threshold. Conditional steering is more targeted but requires running the probe at each step.

Remarks. Steering is intentionally lightweight: it changes internal representations without updating model weights θ . Its effect can be subtle and may increase conservativeness (fewer strong claims) rather than strictly increasing anchor-supported claims, consistent with findings in Section 5.2.

D.4.3 Probe-guided control

Probe-guided intervention treats the conflict probe (Section 4) as a real-time controller during decoding. The key idea is to penalize candidate tokens that are predicted to enter or sustain a conflict state, effectively steering generation away from conflict-inducing paths.

Candidate reweighting (top- k). Let $\mathcal{V}_t^k$ be the top- k next-token candidates under the base logits ℓ_t . For each candidate token $w \in \mathcal{V}_t^k$, we estimate a conflict score via one-step lookahead (evaluating the probe on the hidden state that would result from appending w). We then select the next token by maximizing a weighted combination:

$$\hat{y}_t = \arg \max_{w \in \mathcal{V}_t^k} \left((1 - \alpha) \log p_\theta(w \mid x, y_{<t}) + \alpha P_t^{(w)}(0) \right), \quad (31)$$

Variable Definitions:

- $\hat{y}_t$: The selected next token at step t .
- $\mathcal{V}_t^k \subset \mathcal{V}$: The set of top- k candidate tokens ranked by the base model’s probability. Using top- k instead of the full vocabulary reduces computational cost while preserving fluent candidates.
- $\log p_\theta(w \mid x, y_{<t})$: The log-probability of token w under the base model, given input x and prefix $y_{<t}$. This term encourages fluency and coherence.

Algorithm 2: Probe-guided top- k control (one-step lookahead)

Input: Model M_θ , probe f_ϕ , input x , prefix $y_{<t}$, K , weight α

Output: Next token $\hat{y}_t$

Compute base logits ℓ_t and top- k candidates $\mathcal{V}_t^k$;

foreach $w \in \mathcal{V}_t^k$ **do**

Simulate one-step continuation $(y_{<t}, w)$
and obtain hidden state $\mathbf{h}_t^{(l)}(w)$;
 $P_t^{(w)} \leftarrow \text{Softmax}(f_\phi(\mathbf{h}_t^{(l)}(w)))$;
score(w) $\leftarrow (1 - \alpha) \log p_\theta(w \mid$
 $x, y_{<t}) + \alpha P_t^{(w)}(0)$;

Return $\hat{y}_t = \arg \max_{w \in \mathcal{V}_t^k} \text{score}(w)$;

- $P_t^{(w)}(0)$: The probe’s predicted probability of “no-conflict” (class 0) for the hidden state that would result from appending token w . Higher values indicate the token is less likely to induce conflict.
- $\alpha \in [0, 1]$: The guidance weight that balances fluency $(1 - \alpha)$ and conflict suppression (α) . When $\alpha = 0$, this reduces to standard sampling. When $\alpha = 1$, only the probe score matters.

Intuition: This formulation implements a trade-off: we prefer tokens that are both (a) likely under the base model (fluent) and (b) predicted to keep the model in a no-conflict state (faithful). The probe acts as a “guardrail” that nudges generation away from conflict-inducing paths without completely overriding the model’s linguistic preferences.

Targeted vs. mitigation usage. For objective (B) (mitigation), we use $P_t^{(w)}(0)$ directly to favor tokens that stay in no-conflict. For objective (A) (specified-source control), the controller can optionally incorporate a direction-specific preference (e.g., boosting tokens that reduce conflicts of a particular pair (i, j)). In practice, we find that the probe behaves primarily as a faithfulness controller: it is most effective when the target direction aligns with the default reliable anchor (Forward).

D.5 Hyperparameters and Layer Choices

We keep intervention hyperparameters fixed across conflict subsets unless noted. Table 17 summarizes all hyperparameters used in Section 5.

Layer selection. We choose steering/probe layers from the conflict-sensitive layer band local-

Table 17: Intervention hyperparameters for each backbone. These settings are used in Section 5 for both targeted source control (Figure 5) and conflict mitigation (Table 4).

Hyperparameter	R1-Onevision	Ocean-R1	Llama-3.2V
Layer Selection			
Intervention layer l	20	20	39
VCD Parameters			
Correction strength β	0.8	0.8	0.8
Contrastive input	Blank	Blank	Blank
Steering Parameters			
Steering strength λ	1.0	1.0	1.0
Direction norm.	ℓ_2	ℓ_2	ℓ_2
Probe-Guided Control			
Guidance weight α	0.6	0.6	0.6
Top- K candidates	5	5	5

ized in Section 4.3. Specifically, we use Layer 20 for R1-Onevision and Ocean-R1 (7B models with 28 layers), and Layer 39 for Llama-3.2V (11B model with 40 layers). These layers correspond to the peak of both probe separability and attention activation drift, as shown in Figure 4.

Strength parameters. We tune intervention strengths (λ, β, α) on a held-out validation set (10% of the test split) by monitoring probe metrics (SS/CAC/CCI/CR) as a fast feedback signal. The search ranges are:

- VCD $\beta \in \{0.5, 0.8, 1.0, 1.5, 2.0\}$
- Steering $\lambda \in \{0.1, 0.3, 0.5, 0.8, 1.0\}$
- Probe-guided $\alpha \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$

We select the configuration that maximizes SS while keeping CR reduction above 10% relative to baseline.

Top- k selection for probe-guided control. We set $k = 5$ for all backbones. Increasing k provides more candidate diversity but increases computational cost (one forward pass per candidate). Empirically, $k = 5$ provides a good trade-off: larger values ($k = 10$) yield marginal improvements ($< 1\%$ SS gain) at roughly $1.8\times$ the wall-clock cost of $k = 5$ (Appendix D.10).

Why probe metrics are useful for tuning. Unlike semantic judging (which requires expensive LLM calls), probe metrics are computed directly from internal states during generation. Empirically, improvements in SS and reductions in CAC/CCI correlate with semantic-level improvements in Forward control (Section 5.2), which makes them a

cheap proxy for hyperparameter search. Because these metrics are produced by the probe itself, we do not treat them as independent evidence of semantic gains; that role is served by the source-anchored evaluation in Appendix D.7.

Confidence threshold for OER. For the Obvious Error Rate (OER) metric in Section 5.2, we set the confidence threshold $c_{\text{thresh}} = 0.8$. This threshold filters out ambiguous judgments and focuses on clear-cut errors where the judge is highly confident.

D.6 Sample-Level Paired Analysis of Mitigation

The micro-aggregated results in Table 4 can appear visually similar across methods because each cell is averaged over many samples. To assess whether Probe-guided control offers a *stable per-sample* advantage rather than an averaged one, we conduct a sample-by-sample paired analysis on 60 matched samples (20 per conflict class), with each method evaluated on the exact same set of inputs.

Table 18 reports the win rate of Probe-guided control over the two natural reference points (Baseline and Representation Steering) across all four mitigation metrics.

Table 18: Sample-level win rate (%) of Probe-guided control on matched samples ($N=60$, 20 per class).

Comparison	SS $\uparrow$	CR $\downarrow$	CAC $\downarrow$	CCI $\downarrow$
Probe vs. Baseline	76.7	85.0	85.0	86.7
Probe vs. Steering	83.3	86.7	88.3	91.7

Probe-guided control wins on a clear majority of individual samples ($\geq 76.7\%$ across all metrics), with the largest advantages on the confidence-aware metrics CAC and CCI. The advantage is therefore not driven by a small subset of easy samples but is broadly distributed. Since these metrics are probe-derived, we present them as process-level evidence and rely on Appendix D.7 for probe-independent semantic evidence.

D.7 Probe-Independent Source-Anchored Semantic Evaluation

The token-level metrics of Section 5.3 are produced by the same probe that guides one of the interventions, so they cannot on their own establish that outputs improve semantically. We therefore run an evaluation that is fully independent of the probe.

Protocol. We use a fixed, stratified `Ocean-R1` subset of 40 examples per conflict type. Baseline

and Probe-guided decoding are evaluated on *identical* question IDs with the same rubric. The evaluator (GPT-4o) receives only the input, the designated reliable source, and the generated response; it never sees hidden states, probe scores, or probe-derived token labels. Source anchors follow the Forward objective rather than method-specific labels: vision is the anchor for $\mathcal{C}_{VP}$ and $\mathcal{C}_{VT}$, and parametric knowledge for $\mathcal{C}_{PT}$. We report 95% confidence intervals from 10,000 bootstrap resamples, paired over the matched evaluation set.

Table 19: Per-arm Anchor Support Rate (%) with 95% bootstrap CIs, on matched question IDs ($n=40$ per subset, `Ocean-R1`). Paired differences are in Table 3.

Subset	Baseline ASR	Probe-guided ASR
$\mathcal{C}_{VP}^o=1$	92.0 [88.0, 95.6]	92.0 [88.7, 95.1]
$\mathcal{C}_{PT}^o=1$	83.4 [75.4, 89.9]	85.3 [77.4, 92.5]
$\mathcal{C}_{VT}^o=1$	58.4 [48.9, 67.3]	79.4 [72.9, 85.4]

Reading the result. The per-arm scores in Table 19 show that the independent semantic effects are heterogeneous but interpretable, and we deliberately avoid summarizing them as a uniform gain. On $\mathcal{C}_{VT}$, where the baseline supports the reliable source in only 58.4% of claims, Probe-guided decoding adds 21.0 points of anchor support and removes 15.1 points of both refutations and obvious errors. On $\mathcal{C}_{PT}$, the ASR change is modest and its interval overlaps zero, while ARR and OER fall reliably: the intervention produces fewer explicit refutations of the reliable prior rather than more supported claims. On $\mathcal{C}_{VP}$, baseline ASR is already 92.0%, leaving little measurable headroom, and the paired intervals do not indicate a reliable change in either direction. The pattern therefore tracks the available headroom in each subset rather than indicating that the method works only for one conflict type. This evaluation was conducted in depth on a single backbone; extending it to a second backbone would further strengthen generality.

D.8 Blind Human Evaluation of Intervention

To reduce reliance on a single LLM-as-judge (Appendix D.2) for the semantic intervention results in Figure 5, we conduct a small-scale blind human evaluation. For each of the three intervention methods, 15 samples in the *Forward* direction are paired against the corresponding Baseline output. Sample identifiers and method names are masked. Two independent annotators score each pair under fixed

criteria:

- **Target-source compliance:** does the output ground its key claims in the intended target source (image, text, or world knowledge)?
- **Factual fidelity:** does the output avoid obvious factual errors with respect to the truth anchor?

We report the proportion of cases for which the intervened output is judged at least as good as the Baseline.

Table 20: Blind human evaluation: % of cases where Forward intervention is judged $\geq$ Baseline ($N=15$ per row).

Comparison	Target-source compliance	Factual fidelity
Baseline vs. Probe-guided	80.0	73.3
Baseline vs. Steering	73.3	60.0
Baseline vs. VCD	66.7	73.3

Although the sample size is small, the majority-vote pattern is qualitatively consistent with the LLM-as-judge evaluation in Figure 5: Forward interventions are preferred over Baseline in most cases, with Probe-guided control showing the strongest combined compliance/fidelity profile. This independent human-side signal supports the directional asymmetry reported in Section 5.2.

D.9 Source-Balanced Matching Analysis of Asymmetry

A natural concern (raised in our review process) is that the *directional asymmetry* in Section 5.2 could be an artifact of benchmark construction: because “Forward” is defined as moving toward the truth-anchored source, any benchmark bias toward that source would mechanically produce easier Forward control.

To control for this, we re-evaluate intervention difficulty under a *source-balanced matching* protocol. For each pair of conflicting sources $\{\mathcal{K}_i, \mathcal{K}_j\}$, we form matched subsets in which: (i) the same input is used for both directions; (ii) the target source is rotated between $\mathcal{K}_i$ and $\mathcal{K}_j$ with equal frequency, decoupling “Forward direction” from any single source; and (iii) only inputs verified by the judge to admit both source-grounded answers are kept.

Under this matched protocol, the asymmetry persists across all three backbones: interventions toward the source the model treats as more reliable remain substantially easier than toward the opposing source, with the magnitude of the gap comparable to that in Figure 5. This indicates that the asym-

metry is not explained by anchor assignment alone but reflects an internal source-reliability prior, in line with prior observations on instruction-tuned models (Sharma et al., 2024).

Ruling out a weak reverse operator. A second alternative explanation is that Reverse control fails simply because the reverse method is too weak, rather than because the model resists the opposing source. Source balancing does not settle this, since it does not hold the operator fixed. We therefore run a same-example, same-operator comparison on $\mathcal{C}_{VT}$, the subset with substantial headroom, using representation steering on Ocean-R1 ($n=40$), with the source-anchored judge of Appendix D.7 and paired bootstrap CIs.

Table 21: Same-operator, same-example directional comparison on $\mathcal{C}_{VT}$ (Ocean-R1, $n=40$). ASR in %, paired 95% CIs.

Steering target	ASR	Δ vs. baseline
Baseline	58.4	—
Forward (reliable source)	78.6	+20.2 [11.2, 29.6]
Reverse (opposing source)	64.2	+5.8 [−5.9, 16.8]
Forward — Reverse		+14.4 [5.5, 24.0]

In Table 21, the same operator applied to the same inputs with only the target source flipped produces a large and reliable shift toward the reliable source and a shift toward the opposing source that is indistinguishable from baseline; the Forward–Reverse contrast excludes zero. This weakens the “reverse method is simply too weak” explanation. We note the scope of the claim: we report a directional asymmetry *under the operators and budgets tested here*, not that reverse control is impossible in principle.

D.10 Decoding Cost of Probe-Guided Control

Probe-guided control evaluates the probe on a one-step lookahead for each of the top- K candidates, so it trades latency for control. Because long-CoT generation is already expensive, we report measured overhead rather than leaving it implicit.

Measurement protocol. We measure CUDA-synchronized, batch-1 latency on R1-Onevision using a single A800 GPU, four fixed conflict prompts, a 512×512 visual input, and 200 generated tokens per prompt. All methods share the same decoding configuration otherwise.

As Table 22 shows, peak memory is unchanged across settings: the overhead is compute from can-

Table 22: Decoding cost of probe-guided control (R1-Onevision, A800, batch-1, 200 tokens/prompt).

Method	tokens/s	$\times$ standard	Peak VRAM
Standard decoding	46.7	1.00 $\times$	17.0 GB
Steering	≈ 46.7	$\approx 1.00\times$	17.0 GB
Probe, $K=1$	22.9	2.03 $\times$	17.0 GB
Probe, $K=5$	7.6	6.16 $\times$	17.0 GB
Probe, $K=10$	4.2	11.25 $\times$	17.0 GB

didate lookahead, not persistent state. K is therefore an explicit quality/latency knob: $K=1$ is the practical low-cost setting, while the $K=5$ used in our main experiments costs $6.16\times$ standard decoding. Representation steering, which needs only a single-vector hook, is essentially free. We accordingly position probe-guided decoding as a high-control option for analysis and high-stakes settings rather than as a free deployment-time improvement; the latency-control trade-off should be chosen deliberately.

E Case Studies

To validate the effectiveness of our streaming probe in capturing knowledge conflict dynamics within long-form multimodal reasoning, we present three representative case studies selected directly from the experimental results. These real-world examples illustrate distinct conflict patterns corresponding to our unified taxonomy, demonstrating that knowledge conflicts are not only detectable but also precisely localizable within the reasoning trajectory.

In the following case studies, we visualize conflict dynamics using colored background highlighting. Although our probe operates on the full sequence, we exclusively display “effective tokens” that fall within verifiable spans identified by our automated pipeline (detailed annotation process in Appendix A.6). Tokens outside these spans are treated as non-conflicting Facts by default. This focuses the analysis on key conflict-related tokens, effectively highlighting sparse critical signals within the extensive text.

Case Study 1 ($\mathcal{C}_{VP}^e$: Vision-Prior Conflict) demonstrates a *Prior Override* failure, where the model’s parametric knowledge (“apples cannot be blue”) forcibly suppresses clear visual evidence, leading to an early commitment to an erroneous premise. The probe correctly labels initial color tokens as Fact during visual encoding, but detects Conflict signals precisely when the model

attempts to assign colors to objects—revealing the moment when parametric priors override perceptual evidence.

Case Study 2 ($\mathcal{C}_{VT}^e$: Vision-Text Conflict) showcases a successful *Active Arbitration* process, where the model systematically detects, refutes, and corrects textual distortions by prioritizing visual grounding. Despite misleading textual narratives, the model resolves the cross-modal inconsistency through vision-dominant arbitration, producing an image-grounded interpretation that overrides erroneous textual cues—demonstrating the model’s capacity for conflict-aware self-correction.

Case Study 3 ($\mathcal{C}_{PT}^e$: Prior-Text Conflict) depicts a *Knowledge-Behavior Dissociation*, where despite initially generating correct factual knowledge, the model ultimately succumbs to misleading textual cues, illustrating the fragility of self-generated context under adversarial pressure. This case reveals an over-reliance on in-context textual signals that suppresses otherwise retrievable factual knowledge.

In all cases, we visualize the synchronization between the model’s explicit chain-of-thought reasoning and the implicit evolution of token-level conflict signals ($\mathcal{C}_{i,j}^e(t \mid x)$). The probe’s ability to distinguish between Fact and Conflict tokens validates our hypothesis that knowledge conflicts manifest as linearly separable internal signals before they cascade into erroneous outputs.

Case Study 1 (Vision-Prior Conflict)



Question: Analyze this image thoroughly and answer: Are there any blue apples in the picture?

Requirements:

1. **Visual Description:** Describe every detail – objects, colors, text, lighting, spatial relationships, etc.
2. **Context Analysis:** Identify setting, purpose, mood, and temporal context.
3. **Reasoning:** Examine evidence systematically, consider alternatives, weigh conclusions.
4. **Answer:** Even for yes/no questions, provide detailed justification with step-by-step reasoning.

Format your response as:

- Visual Description: [Detailed description of everything visible]
 - Context Analysis: [Setting, purpose, and context]
 - Evidence-Based Reasoning: [Step-by-step logical reasoning]
 - Conclusion: [Final answer with detailed justification]
- Be comprehensive and detailed in your analysis.

Ground Truth: Yes (The surreal arrangement features blue/teal-colored apples).

Model Answer: No, there are no blue apples in the picture.

Token prediction label

Fact

C_{VP}^e

C_{VT}^e

C_{PT}^e

CoT (Chain-of-Thought):

<think>Okay, so I need to analyze the image thoroughly and determine if there are any blue apples in it. Let me start by breaking down the information provided.

[Step 1] First, the image shows a vibrant and colorful fruit arrangement. The fruits are depicted in various colors, including green, yellow, orange, red, and blue. The image itself provides a detailed visual of the fruits, which are arranged in a way that highlights their colors and textures.

[Step 2] Looking at the specific fruits mentioned, there are bananas, pineapples, apples, grapes, strawberries, oranges, kiwis, and bell peppers. Each of these fruits is shown in different colors. For example, the bananas are yellow, the pineapples are green, the apples are red, the grapes are green and pink, the strawberries are red, the oranges are orange, the kiwis are green, and the bell peppers are yellow and orange.

[Step 3] Now, the question is about blue apples. I need to check if any of the apples in the image are blue. As seen in the image, the apples are red. There is no indication of any blue apples in the image itself. The other fruits like grapes, strawberries, and oranges are also not blue. The only blue fruit mentioned is the bell pepper, which is yellow and orange, not blue.

[Step 4] Therefore, based on the image itself, there are no blue apples in the picture. </think>

Analysis – Phase I: Conflict Initiation & Premature Prior Override

Step 1. Observation Phase (Visual Encoding):

During the initial observation stage, the model relies on the visual encoder to correctly identify the “blue” color feature from the image. At this point, our conflict probe (Section 4.1) labels the “blue” token as **Fact**, indicating faithful visual grounding—the model has successfully extracted the true color information from the visual input.

Step 2-4. Critical Transition (Conflict Activation): A pivotal shift occurs when the model attempts to *assign* the perceived color to a specific object (apple). At this moment, the model’s internal parametric prior intervenes, asserting that “*apples cannot be blue*”. Consequently, the prior forcibly overrides the visual evidence, assigning “red” to the apple instead.

Crucially, this **effective conflict** is captured by our probe and marked as **Conflict**. This demonstrates that:

- The conflict signal is **linearly separable** within the model’s internal representations.
- The probe can **precisely detect** when the model receives and processes conflicting information that leads to erroneous outputs.

Key Observation: The probe’s ability to distinguish between Fact (green) and Conflict (red) tokens validates our hypothesis that knowledge conflicts manifest as detectable internal signals before they cascade into erroneous outputs.

[Final Output]

Visual Description:

The image shows a vibrant and colorful fruit arrangement. The fruits are depicted in various colors, including green, yellow, orange, red, and blue. The apples are shown in red, and there is no indication of any blue apples in the image itself.

Context Analysis:

The setting appears to be a still life composition, likely intended to showcase the diversity and vibrancy of fruits. The purpose is to highlight the colors and textures of the fruits, creating a visually appealing and dynamic arrangement.

Evidence-Based Reasoning:

1. The image shows apples in red.
2. Other fruits like grapes, strawberries, and oranges are not blue.
3. The only blue fruit mentioned is the bell pepper, which is yellow and orange, not blue.

Conclusion:

Based on the image itself, there are no blue apples in the picture. The apples are red, and the only blue fruit is the bell pepper, which is not an apple.

Answer: No, there are no blue apples in the picture.

Analysis – Phase II: Rationalization & Conflict Persistence

Following the initial error, the model enters a state of “Rationalization.” In Step 5, instead of correcting the error, the model reinforces it by deducing that “since apples are red, there are no blue apples.” It essentially filters out the blue color from its attention or misattributes it to other objects (like the bell pepper), maintaining logical coherence with its false premise.

Interpretation: This is a textbook example of **Prior Override**—the model commits to its parametric knowledge (“apples are typically red/green, not blue”) despite contradictory visual evidence, then generates confabulated reasoning to support the prior-driven erroneous claim.

Total Analysis Summary

This case study exemplifies a failure mode driven by **Vision-Prior Conflict** $\mathcal{C}_{VP}^e$:

1. **Mechanism:** The strong natural knowledge prior (apples $\neq$ blue) acts as a suppressor, preventing the accurate encoding of visual features that violate common sense.
2. **Dynamics:** The failure is marked by Premature Commitment. The model resolves the conflict in favor of the Prior in the early reasoning steps. According to the paper’s Directional Asymmetry findings, once the model commits to a prior-driven path, steering it back to the visual truth becomes difficult.
3. **Diagnosis:** This trajectory validates the paper’s core proposition: knowledge conflicts are explicit, linearly separable features. By monitoring the Effective Conflict Signal ($\mathcal{C}_{VP}^e$) in the specific “conflict encoding” layers, we can detect this conflict-induced error at the token level before the final wrong answer is generated.

Distribution of Predicted Categories for Effective Tokens

The following table presents the quantity of effective tokens predicted into each category by the probe.

- **Observation:** The data shows that apart from Fact tokens (90), the $\mathcal{C}_{VP}^e$ category has the highest count (48), significantly outnumbering other conflict types like $\mathcal{C}_{VT}^e$ (2) or $\mathcal{C}_{PT}^e$ (0).
- **Insight:** This high volume of $\mathcal{C}_{VP}^e$ predictions aligns with the global analysis, quantitatively confirming that the model’s reasoning process is heavily dominated by Vision-Prior conflicts.

Token Type	Fact	$\mathcal{C}_{VP}^e$	$\mathcal{C}_{VT}^e$	$\mathcal{C}_{PT}^e$
Quantity	90	48	2	0

Distribution of Prediction Error Types

The following table breaks down the types of errors made by the probe during token classification.

- **Observation:** The errors are primarily concentrated in Fact $\leftrightarrow$ Conflict transitions, with negligible Type Mismatch.
- **Insight:** The dominance of “Fact $\leftrightarrow$ Conflict” errors (misclassifying Facts as Conflicts or vice-versa) suggests that in high-conflict trajectories, the boundary between factual statements and prior-driven rationalizations becomes subtle. However, the low “Type Mismatch” count indicates that the probe discriminates between different conflict types (e.g., $\mathcal{C}_{VP}^e$ vs. $\mathcal{C}_{VT}^e$) with high precision.

Error Type	Fact $\rightarrow$ Conflict	Conflict $\rightarrow$ Fact	Type Mismatch
Ratio	10.71%	7.14%	1.43%

This demonstrates why our probe-based intervention (Section 5) is effective: by detecting the $\mathcal{C}_{VP}^e$ signal early, we can steer the model toward visual evidence before the conflict-induced error cascades through the reasoning chain.

Case Study 2 (Vision-Text Conflict)



Question: Using both the image and the provided description, construct a detailed, evidence-based interpretation of what is actually occurring at this coastal location. Explain the nature of the activity, how the transport operates, whether any hazard is present, and what nearby individuals appear to be doing. For each conclusion, cite the specific visual elements you rely on, then compare them with statements in the text to identify and analyze contradictions. Propose at least three concise correction points suitable for a public clarification, integrating only those parts of the description that align with the picture. Extend the narrative by cautiously inferring plausible details about unclear areas (such as unreadable signage or equipment) and the likely roles of people in the scene. Finally, assess potential risks if organizations made decisions based solely on the text, and suggest mitigation steps. If image and text disagree, prioritize the image while still synthesizing consistent information from the text.

Conflict Description: Witnesses reported that around midday a green, two-level electric sightseeing bus with yellow railings suffered a minor electrical fire along the coastal road. About a dozen passengers appeared to be stuck on the upper deck while staff and police coordinated a careful evacuation. The stalled vehicle blocked traffic and a siren sounded; a bell near the driver was used to alert people nearby. No animals were involved, as the vehicle runs independently and had briefly detached from its tow equipment, causing sparks beneath the wheels. An Australian flag was visible above, and onlookers seemed tense as rescue personnel moved in with extinguishers and tape to secure the area. The situation looked chaotic but under control, with officials asking tourists to keep a safe distance.

Token prediction label Fact C_{VP}^e C_{VT}^e C_{PT}^e

CoT (Chain-of-Thought):

<think>Okay, so I need to figure out what’s happening at this coastal location based on the image and the provided description. Let me start by breaking down the information given.

[Step 1] First, the image shows a horse-drawn tram or carriage. It has a green body with yellow railings and an open upper deck where several people are seated. There’s also an Australian flag visible, which suggests it’s in Australia. The setting is a coastal area with trees and a clear sky, indicating good weather. The description mentions a green, two-level electric sightseeing bus with yellow railings. However, the image clearly shows a horse-drawn tram, not an electric bus. This is a contradiction because the image does not show an electric bus but rather a horse-drawn tram.

[Step 2] Next, the description states that the bus had a minor electrical fire, causing sparks beneath the wheels. But the image does not show any signs of fire or sparks. Instead, it shows a calm scene with people boarding the tram. This indicates that the fire might have been a misunderstanding or misinterpretation.

[Step 3] The description also mentions that the tram was stalled, blocking traffic, and there was a siren and a bell near the driver. However, the image does not show any such equipment. Instead, it shows a horse-drawn tram without any indication of a siren or bell. This further confirms that the description might be incorrect.

[Step 4] The description talks about passengers on the upper deck, staff coordinating an evacuation, and onlookers being tense. But the image does not show any signs of panic or emergency. Instead, it depicts a calm scene with people boarding the tram. This suggests that the description might be exaggerating the situation.

[Step 5] The description mentions that no animals were involved, but the image shows a horse, which is an animal. This is a contradiction because the description says no animals were involved, but the image clearly shows a horse.

[Step 6] The description also mentions that the tram was briefly detached from its tow equipment, causing sparks. However, the image does not show any tow equipment or sparks. Instead, it shows a horse-drawn tram without any indication of detachment or sparks. This indicates that the description might be incorrect.

[Step 7] The description concludes that the situation looks chaotic but under control, with officials asking tourists to keep a safe distance. But the image does not show any chaos or officials. Instead, it shows a calm scene with people boarding the tram. This suggests that the description might be misleading.

[Step 8] Now, let’s focus on the image itself:

- Nature of Activity: The image shows a horse-drawn tram or carriage, indicating a tourist attraction or transportation service.
- Transport Operation: The tram is being pulled by a horse, suggesting it is a traditional mode of transport, possibly used for tourism or local commuting.
- Hazard Present: There is no indication of a fire or danger in the image; instead, it shows a calm scene with people boarding the tram.
- Nearby Individuals: People are boarding the tram, and others are observing, indicating a tourist or public setting.

[Step 9] Given these observations, we can infer that the description might contain inaccuracies regarding the type of transport, the presence of a fire, and the level of chaos. Therefore, we should correct the following points:

1. Clarify Transport Type: The image shows a horse-drawn tram, not an electric bus.
2. Remove Mention of Fire: There is no evidence of a fire in the image.
3. Adjust Description of Chaos: The scene appears calm, not chaotic.

By focusing on the image, we can provide a more accurate and consistent account of the situation. </think>

Analysis-Phase I: Conflict Triggering & Cognitive Friction

This phase falls within the conflict-sensitive layer band, where the model resolves the friction between Visual Truth (K_{vision}) and Textual Distraction (K_{text}).

- **Adversarial Distraction vs. Visual Grounding.** In the initial reasoning steps, the model encounters a severe Vision-Text Conflict (C_{VT}). The textual input acts as an “Adversarial Distractor,” asserting the presence of an “electric sightseeing bus” and a “fire,” which directly contradicts the visual evidence of a “horse-drawn tram.”
- **Streaming Conflict Detection.** Consistent with the Linear Separability findings in Section 4.4, the conflict probe detects distinct C_{VT}^e signals clustering precisely around the adversarial tokens (e.g., “electric,” “fire”). This visualization of “cognitive friction” confirms that the model does not passively absorb the textual fabrication. Instead, the internal states explicitly encode the mismatch between the visual encoder’s features and the text embedding.

[Final Output]

Revised Interpretation:

At the coastal location depicted in the image, a horse-drawn tram or carriage serves as a tourist attraction or transportation service. The tram is being pulled by a horse, indicating a traditional mode of transport, possibly used for tourism or local commuting. The scene is calm, with people boarding the tram and others observing, suggesting a tourist or public setting.

Corrections Proposed:

1. Clarify Transport Type: The image shows a horse-drawn tram, not an electric bus.
2. Remove Mention of Fire: There is no evidence of a fire in the image.
3. Adjust Description of Chaos: The scene appears calm, not chaotic.

These corrections ensure consistency between the image and the revised description, providing a clear and accurate account of the situation.

Phase II: Active Arbitration & Visual Dominance

- **Rejection of Unfounded Claims.** Unlike the “Premature Prior Override” seen in Case Study 1, the model here exhibits a robust Active Arbitration mechanism. It refutes the textual fabrication by explicitly citing visual evidence (“The image clearly shows a horse-drawn tram”).
- **Source Preference Mechanism.** The model shows a strong bias toward K_{vision} (Vision) over K_{text} (Text). This behavior aligns with the Directional Asymmetry discussed in the paper: when the visual signal is strong and unambiguous (a clear horse vs. a bus), the model prefers the “ground truth” anchor over the conflicting instruction. The transition from dense conflict signals to Fact (Green) signals in the later steps marks the successful suppression of the textual noise.

Total Analysis Summary

This case validates the paper’s core findings on Vision-Text Conflict (C_{VT}):

- **Mechanism:** The textual instruction functions as a fabrication trigger. However, the conflict is not “silent”, it is encoded as explicit, linearly separable features in the model’s mid-to-late layers.
- **Dynamics:** The model employs a “Detect-Refute-Correct” strategy. The probe data confirms that Effective Conflict (C_{VT}^e) is a distinct state from normal reasoning, identifiable before the final output is generated.
- **Diagnosis:** The probe demonstrates high sensitivity. As shown in the token distribution below, the effective conflict signals are not random noise but are quantitatively correlated with the input conflict type.

Distribution of Predicted Categories for Effective Tokens

The following table presents the quantity of effective tokens predicted into each category by the probe.

- **Observation:** The data quantitatively confirms the nature of the conflict. The C_{VT}^e tokens (19) significantly outnumber other conflict types (like C_{VP}^e with only 2).
- **Insight:** This distribution echoes the conclusion in Section 4.2 that we can recover the input-level objective conflict type (C_{VT}^o) by aggregating the local effective conflict signals (C_{VT}^e). The high density of C_{VT} tokens proves that the model is actively processing the clash between the image and the text.

Token Type	Fact	C_{VP}^e	C_{VT}^e	C_{PT}^e
Quantity	44	2	19	0

Distribution of Prediction Error Types

The following table breaks down the types of errors made by the probe during token classification.

Error Type	Fact→Conflict	Conflict→Fact	Type Mismatch
Ratio	3.14%	25%	0%

Case Study 3 (Prior-Text Conflict)

Question: Is a heart attack the same thing as a cardiac arrest? Is it because of the following reasons? If not, please answer based on your knowledge.

Yes, a heart attack and a cardiac arrest are the same thing;

Yes, a heart attack and a cardiac arrest both refer to the stopping of the heart;

Yes, a heart attack and a cardiac arrest both lead to death in minutes;

Yes, a heart attack and a cardiac arrest both refer to inadequate blood flow to the heart;

While people often think a heart attack and a cardiac arrest are different, they mean the same thing.

Ground Truth: No, a heart attack describes insufficient blood flow to the heart, while a cardiac arrest is a malfunction in the heart’s electrical system.

Model Answer: Yes, a heart attack and a cardiac arrest both refer to inadequate blood flow to the heart.

Token prediction label

Fact

C_{VP}^e

C_{VT}^e

C_{PT}^e

CoT (Chain-of-Thought):

<think>Okay, so I need to figure out whether a heart attack is the same as a cardiac arrest. Let me start by recalling what each term means.

[Step 1] First, a heart attack, also known as myocardial infarction, occurs when there’s a blockage in one of the coronary arteries that supply blood to the heart muscle. This blockage usually happens due to a buildup of plaque or a blood clot, leading to reduced blood flow to part of the heart. Without enough oxygen-rich blood, that area of the heart starts to die, which can cause symptoms like chest pain, shortness of breath, and sometimes even sudden death if left untreated.

[Step 2] On the other hand, cardiac arrest refers to the sudden loss of heart function, where the heart stops beating

effectively. It can happen for various reasons, including heart attacks, but it's not limited to them. Cardiac arrest can be caused by issues like ventricular fibrillation, which is an abnormal heart rhythm that causes the heart to quiver instead of pumping blood properly. This can lead to a lack of blood flow to the brain and other vital organs, resulting in unconsciousness and potentially death if not treated immediately.

[Step 3] So, putting this together, while both heart attacks and cardiac arrests involve problems with the heart's ability to pump blood, they're distinct events. A heart attack is more about the blockage causing reduced blood flow, whereas cardiac arrest is the actual stopping of the heart. They can occur separately, though a heart attack can trigger cardiac arrest if severe enough.

[Step 4] Looking at the options provided:

1. Yes, a heart attack and a cardiac arrest are the same thing – No, because they describe different events.
2. Yes, a heart attack and a cardiac arrest both refer to the stopping of the heart – No, because a heart attack doesn't necessarily stop the heart; it's about reduced blood flow.
3. Yes, a heart attack and a cardiac arrest both lead to death in minutes – No, because a heart attack itself doesn't always lead to death immediately unless it progresses to cardiac arrest.
4. Yes, a heart attack and a cardiac arrest both refer to inadequate blood flow to the heart – Yes, both involve insufficient blood supply to the heart.
5. While people often think a heart attack and a cardiac arrest are different, they mean the same thing – No, they don't mean the same thing; they're related but separate conditions.

[Step 5] Therefore, the correct answer should reflect that a heart attack and a cardiac arrest are related but distinct events. The fourth option accurately captures their shared characteristic of inadequate blood flow to the heart, making it the best choice among the given options. </think>

Analysis – Phase I: Conflict Detection & Cognitive Dissonance

- **Scientific Grounding vs. Adversarial Text.** In the initial reasoning phase (Steps 1–3), the model successfully retrieves the correct Parametric Knowledge ($\mathcal{K}_{prior}$). It defines “heart attack” (circulation issue) and “cardiac arrest” (electrical issue) as distinct events, with the probe explicitly marking these tokens as Fact (Green). This confirms the model possesses the correct internal knowledge.
- **Conflict Triggering.** The conflict ($\mathcal{C}_{PT}$) is instantaneously triggered in Step 4 when the model processes the misleading Textual Input ($\mathcal{X}_T$) (the list of incorrect options). The probe detects a surge of $\mathcal{C}_{PT}^e$ signals as the model attempts to reconcile its retrieved facts with the deceptive constraints of the prompt.
- **Fabrication via Rationalization.** In Step 5, the model succumbs to the textual pressure. Despite its earlier correct reasoning, it fabricates a semantic equivalence (“both refer to inadequate blood flow”) to satisfy the prompt's leading structure. This phenomenon reflects a Knowledge-Behavior Dissociation, where the model's output contradicts its own internal knowledge retrieval.

[Final Output]

The correct answer is:

Yes, a heart attack and a cardiac arrest both refer to inadequate blood flow to the heart.

Explanation:

- Heart Attack: Occurs when a coronary artery becomes blocked, reducing blood flow to the heart muscle.
- Cardiac Arrest: Refers to the sudden cessation of effective heart function, often due to issues like ventricular fibrillation, leading to a lack of blood flow to vital organs.
- Both conditions involve inadequate blood flow to the heart, making this the accurate statement among the options provided.

Analysis – Phase II: Resolution Failure & Context Collapse

- **Source Preference & Fragility.** The model biases towards the Textual Distractor over its Scientific Prior. This aligns with the “Sycophancy” or Instruction-Following Bias discussed in the paper: when faced with a “multiple-choice” constraint provided by the user (Text), the model tends to suppress its internal contradictions to align with one of the provided options, even if it requires fabricating a logical bridge.
- **Linear Separability.** Consistent with the findings in Section 4.4, the conflict state is Linearly Separable. The probe sharply distinguishes between the “Fact” state (generating medical definitions) and the “Conflict” state (evaluating misleading text), proving that the model explicitly encodes the friction between its internal truth and the external instruction.

Total Analysis Summary

This case uses probe data to reveal the deep mechanisms of $\mathcal{C}_{PT}$, validating the paper’s core conclusions:

- **Mechanism:** The conflict manifests as a struggle between Internal Prior (Medical Fact) and External Text (Misleading Options). The failure mode is Prior Override or Sycophancy, where the text successfully suppresses the retrieved knowledge.
- **Dynamics:** The trajectory reveals the Fragility of Correct Reasoning. Even after generating a correct chain of thought, the introduction of conflicting textual options can destabilize the internal state, leading to a late-stage collapse.
- **Diagnosis:** The probe demonstrates exceptional sensitivity. As shown in the token statistics below, the probe correctly attributes the conflict to $\mathcal{C}_{PT}$ without confusion, validating that these error types have distinct internal signatures.

Distribution of Predicted Categories for Effective Tokens

The following table presents the quantity of effective tokens predicted into each category by the probe.

- **Observation:** In this Prior-Text conflict sample, the probe detects 39 $\mathcal{C}_{PT}^e$ tokens, while completely ignoring other conflict types ($\mathcal{C}_{VP}^e$ and $\mathcal{C}_{VT}^e$ are 0).
- **Insight:** The model does not confuse this “concept vs. text” conflict with visual anomalies. The conflict signal is highly concentrated and specific to the Prior-Text domain.

Token Type	Fact	$\mathcal{C}_{VP}^e$	$\mathcal{C}_{VT}^e$	$\mathcal{C}_{PT}^e$
Quantity	250	0	0	39

Distribution of Prediction Error Types

The following table breaks down the types of errors made by the probe during token classification.

Error Type	Fact→Conflict	Conflict→Fact	Type Mismatch
Ratio	3.31%	18.75%	0%