DO VLMS FAIL AT CAUSAL REASONING OR OUTPUT FORMAT PRODUCTION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluations of vision-language model (VLM) causal reasoning typically rely on a single output format, making it difficult to distinguish failures of format production from failures of reasoning. We introduce a diagnostic framework that combines structure-first generation, multi-format testing across four structural representations, three levels of prompt scaffolding, and score-compliance pairing, where content quality is reported alongside format compliance. Across six VLMs spanning 2023–2025 architectures, we find that the structural expression of causal reasoning is highly sensitive to format and prompt support. The same model reaches 97% compliance for numbered sequences but below 1% for arrows, while removing scaffolding yields compliance drops of 25–91 points. For five of the six models, structural and textual quality converge within the same response once the format barrier is removed. Targeted fine-tuning further helps distinguish limitations arising from insufficient training signal from those that persist under the tested architecture.

1 INTRODUCTION

Vision-language models (VLMs) generate fluent causal explanations when queried about images. Given a photograph of a beach scene and the question "What would happen if strong winds began blowing from the mountains?", a VLM responds coherently with "The umbrellas may topple or fly away." But how well do current evaluations test whether this fluency reflects causal reasoning? Existing benchmarks rely on text-based answers, selection from predefined causal chains, or binary judgments (Chen et al., 2024; Zhang et al., 2025; Komanduri et al., 2025). All three allow success through autoregressive pattern completion without any structural reasoning. None varies prompting support to test whether measured capabilities are self-sustained or depend on how carefully the model is prompted. These two evaluation gaps have deployment consequences. In safety-critical applications such as medical image interpretation (Hartsock & Rasool, 2024) and autonomous driving (Zhou et al., 2024a), causal reasoning cannot rely on curated prompts or worked examples.

Both gaps connect to a distinction in explainable AI between *plausibility* and *faithfulness* (Lyu et al., 2024; Agarwal et al., 2024). Plausible outputs pass as logical through linguistic fluency. Faithful outputs accurately reflect the underlying reasoning. These properties can diverge (Madsen et al., 2024), and VLMs are a case in point. Models generate confident explanations while ignoring contradictory visual evidence (Guan et al., 2024; Tong et al., 2024). They verbalize causal narratives through pattern association, not structured reasoning (Zečević et al., 2023; Chen et al., 2024). Even in selection-based tasks, where models choose among predefined arrow-notation causal chains, performance degrades as reasoning complexity increases (Zhang et al., 2025). None of these evaluations tests whether models can decompose causal reasoning into explicit structural form. A harder test would require the model to generate a causal structure before any textual explanation. Structure-first generation requires the model to specify causal relationships explicitly, without the aid of linguistic fluency. But does success on such a task reflect causal competence, or format familiarity?

A single structural format is insufficient to distinguish these possibilities reliably. A model that generates valid causal chains in arrow notation may be drawing on format fluency acquired during pretraining, not necessarily causal reasoning. If the same model succeeds in one format but fails in another for identical causal content, however, the limitation is more plausibly format-specific than a general absence of causal ability. Failure across several structurally distinct formats provides

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

stronger evidence of a reasoning limitation. Multi-format evaluation can therefore separate these cases more effectively than a single-format test. Success in one format does not establish genuine understanding, nor does failure in one format establish its absence. Variation across formats instead provides a behavioral diagnostic that complements text-based or selection-based measures. Research on human causal cognition similarly supports the view that genuine understanding involves structured internal representations (Friston, 2010; Battaglia et al., 2013).

The multi-format design addresses the first evaluation gap, whether causal competence is present beneath format-specific fluency. The second, whether that competence is self-sufficient or prompt-dependent, requires a separate analysis. Research on large language models (LLMs) has shown output format to be a serious confound, with systematic biases varying across models (Long et al., 2025; Yang et al., 2026). Beyond format, even when a model succeeds under a given format, does that success depend on external scaffolding, such as worked examples, framework references, or detailed instructions? A model that generates valid structures only under full scaffolding exhibits a capability that depends on external support. In deployment settings without carefully designed prompts, such capability may not remain reliable.

To address these two evaluation gaps, we develop a layered diagnostic framework testing how reliably VLM causal reasoning can be measured across formats and scaffolding, and using targeted fine-tuning to diagnose the source of identified fragilities.

Layer 1. **Dual-Probe Evaluation.** For each image and causal question, the model first generates a causal structure in a specified format, then a textual explanation. The structure-first ordering requires the model to specify causal relationships before linguistic fluency can intervene. Divergence between structural and textual quality within the same response is the primary diagnostic signal, whose source the subsequent layers investigate. Causal questions target three levels of Pearl's causal hierarchy (Pearl & Mackenzie, 2018), from associative (observable co-occurrences) through interventional (effects of hypothetical actions) to counterfactual (alternative outcomes under changed conditions).

Layer 2. **Format Robustness.** The same causal content is requested in four structural formats. Three are linearized (arrow notation, semicolon-separated sequences, numbered sequences) and one is graph-level (adjacency lists with parent-child mappings). A model whose structural quality varies across formats shows format-specific fragility.

Layer 3. **Scaffolding Robustness.** Each model-format combination is tested under three levels of prompt-based support (full, partial, and minimal). A model whose structural quality degrades when scaffolding is reduced is scaffolding-dependent. Format-scaffolding combinations where structural quality collapses constitute the model's *blind spots*.

Layer 4. **Targeted Intervention.** For models with identified blind spots, we apply fine-tuning in two stages. Stage 1 (single-format, single-scaffolding) trains on one blind-spot condition and tests whether the improvement is format-robust and scaffolding-robust. Stage 2 (multi-format, multi-scaffolding) trains across all format and scaffolding combinations to determine whether the Stage 1 fragilities are resolvable through training diversity or reflect architectural capacity boundaries.

This structure-first, multi-format diagnostic framework is applied to six VLMs spanning five training approaches across a 2023–2025 capability range. We introduce four diagnostic metrics (Abstraction Gap, Causal Depth Disparity, Format Variance, and Scaffolding Sensitivity Index), a five-category failure mode taxonomy, and score-compliance pairing as the primary unit of analysis. Our contributions are as follows.

- We show that **single-format evaluations conflate format production with causal competence**. The same model under identical conditions yields 97% format compliance for numbered sequences but below 1% for arrows. Models that appear incapable of causal reasoning under one format can generate high-quality causal structures under another. For five of the six tested models, structural and textual quality converge within the same response once the format barrier is removed, a pattern that becomes visible only when scores are considered together with format compliance.
- We show **scaffolding dependence is format-specific, with qualitatively different failure mechanisms across formats**. The same model fails through Cross-Format Substitution for arrows, Ungrounded Structure for semicolons, and near-zero failure for numbered lists, all under identical

minimal scaffolding. Evaluations holding scaffolding fixed cannot detect these format-dependent fragilities.

- We identify **a structural complexity boundary between linearized chains and graph decomposition.** No tested model crosses without scaffolding support. The strongest model maintains near-perfect compliance across three linearized formats but drops to 44% for adjacency lists under minimal scaffolding.
- We show **targeted fine-tuning, used as a controlled intervention, separates training data limitations from architectural constraints**. Single-format training creates a dissociation between format and scaffolding fragility. Multi-format multi-scaffolding training shows for one architecture both fragilities were data bottlenecks, while for another, scaffolding fragility resolves but format capacity is architecturally bounded.

2 Related Work

We situate our work at the intersection of three threads, all connected to the broader question of when benchmark scores reflect genuine capability versus fluency with the output format. Extended discussion is in Appendix A.1.

Faithfulness vs. plausibility. The distinction between faithful and plausible explanations is well established (Jacovi & Goldberg, 2020; Lyu et al., 2024; Agarwal et al., 2024). CoT traces may not faithfully reflect internal processing in LLMs (Lanham et al., 2023; Turpin et al., 2023) or VLMs (Balasubramanian et al., 2025), and unlearning reasoning steps can change predictions (Tutek et al., 2025). Our framework complements these approaches by testing whether the structural expression of causal reasoning holds across formats and scaffolding conditions, and whether targeted fine-tuning can resolve identified fragilities.

Output format as evaluation confound. LLM performance varies dramatically across output formats (Sclar et al., 2024; Tam et al., 2024), with systematic model-specific biases (Long et al., 2025) and low accuracy on structured generation benchmarks (Yang et al., 2026). Our work shows this confound is equally severe for VLM causal reasoning, with format compliance varying from near-ceiling to near-zero within the same model depending on the requested notation.

VLM causal reasoning benchmarks. Existing benchmarks use multiple-choice selection (Chen et al., 2024; Li et al., 2025), counterfactual scenarios (Zhang et al., 2024), causal chain verification (Zhang et al., 2025), temporal inference (Wang et al., 2025), and formal graph tasks on synthetic images (Komanduri et al., 2025). Each uses a single output format and characterizes weaknesses without diagnosing their source. Our work introduces the first multi-format, structure-first evaluation of VLM causal reasoning, and uses targeted fine-tuning to distinguish training artifacts from architectural constraints.

3 Methodology

Models and Data. We evaluate six VLMs spanning five training approaches. *LLaVA-NeXT 13B* (Liu et al., 2024) (synthetic recaptioning), *LLaVA-RLHF* (Sun et al., 2023) (RLHF alignment), *LLaVA-CoT* (Xu et al., 2024) (chain-of-thought supervision), *CogVLM2* (Hong et al., 2024) (deep vision-language fusion via Visual Expert), *MiniGPT-4* (Zhu et al., 2023) (standard projection baseline), and *Qwen3-VL-Instruct 8B* (Bai et al., 2025) (modern interleaved fusion).

We construct two non-overlapping datasets from MS-COCO (Lin et al., 2014). The **evaluation dataset** contains 1,000 images from the COCO validation set (5 captions each). For each image, GPT-4o OpenAI (2024) generates nine causal questions targeting Pearl's three causal levels (Pearl & Mackenzie, 2018), from associative (L1) through interventional (L2) to counterfactual (L3), yielding 9,000 questions total (see Appendix A.4 for an example from our datasets). Recent work has shown LLMs generate high-quality synthetic VQA data when validated against human judgments (Su et al., 2025). We manually verify a random sample of 200 questions. Over 95% are answerable from the image and correctly target their causal level. The **training dataset** contains 5,000 images from the COCO training set with GPT-4o-generated causal questions and ground-truth structures in arrow notation. Structures in the other three formats are obtained via rule-based conversion.

3.1 LAYER 1: DUAL-PROBE EVALUATION

For each model, we collect responses under two probes on the same images and questions.

The **Text-Only Probe** prompts the model to answer a causal question in free text, with no structural output requested. This measures plausible causal language generation.

The **Structure-Text Probe** prompts the model to first generate a causal structure in a specified format, then provide a textual explanation. The structure-first ordering requires the model to specify causal relationships explicitly before linguistic fluency can intervene. Divergence between structural and textual quality within the same response is the primary diagnostic signal, whose source the subsequent layers investigate.

We quantify the cost of requiring structural output with the **Abstraction Gap** ($\text{AG} = S_{\text{TO}} - S_{\text{ST}}^{\text{text}}$), where S_{TO} is the Text-Only score and $S_{\text{ST}}^{\text{text}}$ is the text component of the Structure-Text response. Positive AG indicates requiring structural output degrades the model's textual quality. This connects to prior work showing reasoning performance depends on output representation (Tam et al., 2024).

3.2 LAYER 2: FORMAT ROBUSTNESS

The Structure-Text Probe requests causal structures in four formats. For the chain *Wind* $\rightarrow$ *Force on umbrellas* $\rightarrow$ *Umbrellas topple*, the four representations are as follows.

Arrow:
```
Wind -> Force on umbrellas -> Umbrellas topple
```

Numbered:
```
1. Wind 2. Force on umbrellas 3. Umbrellas topple
```

Semicolon:
```
Wind; Force on umbrellas; Umbrellas topple
```

Adjacency List:
```
Wind: Force on umbrellas
Force on umbrellas: Umbrellas topple
Umbrellas topple: (none)
```

The first three are **linearized chains** differing only in the delimiter. The adjacency list requires decomposing the chain into pairwise parent-child relationships with explicit terminal markers, making it qualitatively different. These four formats span a gradient from patterns deeply embedded in pretraining data (numbered lists) to representations rare in pretraining corpora (adjacency lists).

Standard evaluation metrics do not capture how model behavior varies across structural representations and reasoning conditions. We introduce diagnostic metrics quantifying specific evaluation dimensions (see Appendix A.2 for per-metric justification). Since our evaluation judges score causal content quality regardless of format compliance (§3.5), structural scores alone can be misleading when compliance is low. We therefore treat the score-compliance pair as the primary unit of analysis. The structural score indicates causal content quality, while format compliance indicates whether the model generated the requested format. **Format Variance** (FV) is the standard deviation of structural scores across the four formats for a given model, scaffolding, and causal level. Low FV indicates format-robust competence. High FV indicates format-specific fragility, consistent with prior work on output format sensitivity (Long et al., 2025; Yang et al., 2026). We also compute **Causal Depth Disparity** (CDD), the normalized difference between textual and structural scores within a single response ($\text{CDD} = (S_{\text{ST}}^{\text{text}} - S_{\text{ST}}^{\text{struct}})/S_{\text{ST}}^{\text{text}}$). CDD is interpretable only at high compliance. At low compliance, format failure is the primary signal and CDD adds no information beyond what compliance already provides. Complete CDD values are in Appendix A.8.

3.3 LAYER 3: SCAFFOLDING ROBUSTNESS

Each model-format combination is tested under three levels of prompt-based support. Full templates are in Appendix A.3.

C1 (Full). Three worked examples (from images outside the evaluation set) showing both structural and textual output, Pearl's causal framework reference, and detailed construction instructions.

C2 (Partial). Framework reference and instructions retained. Worked examples removed.

C3 (Minimal). Brief task description with basic format indication only. No examples, no framework reference, no detailed instructions.

This design isolates two scaffolding components. The **Scaffolding Sensitivity Index** and its decomposition quantify the effect of each.

$\text{SSI}_{\text{full}} = S_{\text{C1}}^{\text{struct}} - S_{\text{C3}}^{\text{struct}}$

$\text{SSI}_{\text{ex}} = S_{\text{C1}}^{\text{struct}} - S_{\text{C2}}^{\text{struct}}$ (example removal)

$\text{SSI}_{\text{instr}} = S_{\text{C2}}^{\text{struct}} - S_{\text{C3}}^{\text{struct}}$ (instruction removal)

Positive SSI indicates degradation when scaffolding is removed. Negative SSI indicates improvement. Large positive SSI signals scaffolding dependence, a capability externally conditioned and not self-sufficient. Note that $\text{SSI}_{\text{instr}}$ conflates removal of the Pearl framework reference and the detailed procedural instructions. We do not claim to isolate these individually. Format-scaffolding combinations where a model's structural quality collapses constitute the model's **blind spots**, which motivate the targeted interventions in Layer 4.

3.4 Layer 4: Targeted Intervention

For models with identified blind spots, we apply fine-tuning as a targeted intervention to test whether format and scaffolding fragilities are addressable. We conduct two independent fine-tuning stages, each starting from the base model, using standard cross-entropy training for one epoch (3,000 steps) with identical hyperparameters (Appendix A.5).

Stage 1 (Single-Format, Single-Scaffolding). The model is trained on one blind-spot format at one scaffolding level. We evaluate whether the improvement is *format-robust* (does it transfer to untrained formats?) and *scaffolding-robust* (does it transfer to untrained scaffolding levels?). A dissociation between these two dimensions indicates format production and scaffolding dependence are separable components of structural capability.

Stage 2 (Multi-Format, Multi-Scaffolding). The model is trained on structures spanning all four formats and all three scaffolding levels. The training dataset contains the same 5,000 images and 45,000 question-answer pairs as Stage 1. The L2 and L3 questions (30,000 total) are distributed equally across all twelve format-scaffolding combinations (2,500 per combination), cycling systematically so that every two images cover all combinations. Fragilities diminishing under training diversity are training signal effects. Those persisting point toward architectural capacity boundaries.

Model selection is detailed in §4.4.

3.5 Evaluation Infrastructure

Scoring. Two independent LLM judges (Claude 4.5 Sonnet Anthropic (2025), Gemini 2.5 Flash Comanici et al. (2025)) score outputs from 0–10 for correctness, completeness, and causal validity, independent of format compliance, which is evaluated separately. LLM-based evaluation provides a scalable assessment correlating strongly with human judgments (Gu et al., 2025; Ho et al., 2025). Neither judge shares a model family with any evaluated VLM. Outputs classified as Complete Non-Response or Structure Omission receive automatic structural scores of 0.

Human validation. Expert annotators independently score 6,480 responses (10 images $\times$ 9 questions $\times$ 4 formats $\times$ 3 scaffolding $\times$ 6 VLMs, with text and structure assessed separately at L2 and L3). Pearson correlation between human and LLM judge ratings is 0.84 for structure scores. Human evaluation confirms the central findings, including numbered format scaffolding robustness, the structural complexity boundary between linearized and graph-level formats, and RLHF suppression at full scaffolding (Appendix A.7).

Failure Mode Taxonomy. Every structural output is classified into one of five categories using rule-based automated parsing, independent of the LLM judges. The five categories are *Format-Compliant Structure* (well-formed, image-grounded), *Complete Non-Response* (both text and structure missing), *Structure Omission* (text present, structure absent), *Cross-Format Substitution* (structure in a different format from requested), and *Ungrounded Structure* (correct format but abstract placeholders in place of image-specific content). Classification uses format-specific regex patterns and node-caption matching against the image's COCO captions, with manual verification on 100 responses per category (misclassification $<5\%$).

4 EXPERIMENTS AND RESULTS

Each of six models is evaluated under two probes, four formats, and three scaffolding conditions, yielding 54,000 Text-Only and 648,000 Structure-Text responses. As described in §3.2, we report score-compliance pairs throughout. The score indicates causal content quality regardless of format, and compliance indicates whether the model generated the requested format.

Table 1: Text-Only Probe scores (averaged across judges, 0–10). Per-judge breakdown in Appendix A.6.

Model	L1	L2	L3
LLaVA-NeXT	8.18	8.11	7.97
LLaVA-RLHF	4.03	7.24	7.41
LLaVA-CoT	8.46	8.42	8.07
CogVLM2	8.83	8.29	8.07
MiniGPT-4	6.72	6.89	5.96
Qwen3-VL	8.46	8.35	8.36

4.1 LAYER 1: THE PLAUSIBILITY BASELINE

Table 1 reports Text-Only Probe scores. These establish the plausibility ceiling against which structural output is compared. Most models generate strong causal text. CogVLM2 (8.83), LLaVA-CoT (8.46), and Qwen3-VL (8.46) score above 8.0 at L1, with gradual degradation at higher causal levels consistent with increasing difficulty. MiniGPT-4 generates the weakest text across all levels (6.72, 6.89, 5.96) and serves as the baseline anchor. We return to this model's text quality in §4.4, where structural fine-tuning yields an unexpected text quality improvement. LLaVA-RLHF shows an anomalous L1 inversion (4.03) consistent with RLHF alignment suppression patterns detailed in §4.3.

Table 2: L3 structural scores (averaged across judges, 0–10) with format compliance (%) as subscript. C1/C2/C3 = full/partial/minimal scaffolding. Bold = highest per condition. At low compliance (<50%), scores reflect the format actually generated, not the requested format.

	Arrow			Semicolon			Numbered			Adj. List		
	C1	C2	C3	C1	C2	C3	C1	C2	C3	C1	C2	C3
NeXT	$6.73_{99.5}$	$6.60_{98.5}$	$6.37_{10.3}$	$7.28_{99.7}$	$6.58_{99.2}$	$3.97_{12.5}$	$7.43_{99.7}$	$7.23_{98.9}$	$6.81_{98.3}$	$5.82_{94.0}$	$4.33_{39.3}$	$1.63_{2.7}$
RLHF	$0.46_{8.4}$	$5.70_{65.0}$	$5.95_{4.6}$	$4.55_{72.3}$	$4.80_{76.8}$	$2.52_{14.2}$	$6.25_{85.2}$	$6.50_{96.9}$	$5.69_{87.8}$	$0.00_{0.0}$	$2.36_{2.6}$	$0.57_{0.0}$
CoT	$6.77_{99.5}$	$7.16_{99.3}$	$7.03_{79.9}$	$6.76_{98.5}$	$3.06_{67.9}$	$5.08_{85.1}$	$7.54_{99.4}$	$7.45_{99.3}$	$7.32_{97.3}$	$6.55_{96.1}$	$5.73_{87.9}$	$4.47_{71.6}$
CogV2	$2.85_{48.5}$	$3.25_{13.2}$	$4.46_{33.6}$	$4.16_{66.4}$	$1.73_{20.8}$	$1.59_{13.0}$	$3.97_{55.4}$	$2.81_{45.6}$	$4.46_{84.4}$	$0.99_{11.1}$	$0.94_{6.7}$	$0.92_{13.8}$
Mini	$4.30_{20.5}$	$4.87_{13.1}$	$5.48_{0.6}$	$5.22_{66.5}$	$4.71_{36.1}$	$4.78_{0.8}$	$6.16_{99.3}$	$6.07_{92.8}$	$5.74_{97.7}$	$3.43_{56.5}$	$2.98_{62.2}$	$2.00_{48.4}$
Qwen	$\mathbf{7.79}_{99.2}$	$\mathbf{7.74}_{98.6}$	$\mathbf{7.85}_{97.2}$	$\mathbf{7.83}_{99.4}$	$\mathbf{7.80}_{98.9}$	$\mathbf{7.77}_{97.8}$	$\mathbf{8.04}_{99.1}$	$\mathbf{8.11}_{98.5}$	$\mathbf{8.18}_{97.8}$	$\mathbf{7.55}_{97.9}$	$\mathbf{6.41}_{82.9}$	$\mathbf{4.72}_{44.2}$

4.2 LAYER 2: FORMAT ROBUSTNESS

Table 2 presents the central diagnostic result, L3 structural scores paired with format compliance across all formats and scaffolding conditions.

Causal reasoning is present but not format-robust. MiniGPT-4 at minimal scaffolding illustrates the pattern. Numbered format yields 5.74 with 97.68% compliance, while arrow format yields 5.48 with 0.58% compliance. The causal content quality is comparable, but the format production is not. A single-format evaluation using arrows would conclude this model lacks structural capability. The numbered result shows the content quality is preserved but the format channel is blocked. This pattern holds for five of six tested models. The contrast is sharpest between linearized formats and the adjacency list. At C3, adjacency list scores range from 0.57 to 4.47 while numbered scores range from 4.46 to 7.32 for the same models. Only Qwen3-VL maintains consistent scores across the three linearized formats (7.77–8.18 at C3, all above 97% compliance).

At high compliance, structural quality tracks text quality. When compliance exceeds 85%, within the same response, structural scores approach text scores for five of six tested models. For numbered format at C1, where all models except CogVLM2 exceed 85% compliance, CDD values range from -0.046 to $+0.001$ (Appendix A.8). The measured gap between plausible text and structural output is driven primarily by format production failure, not by a divergence between textual and structural quality. Only CogVLM2 retains a substantial quality gap (CDD = 0.332) even at accessible formats, consistent with its structural weakness across all conditions. Within the tested models,

Table 3: Failure mode distributions (%) at minimal scaffolding (C3). FC = Format-Compliant, NR = Non-Response, SO = Structure Omission, CFS = Cross-Format Substitution, UG = Ungrounded Structure. Dominant failures (>40%) are underlined.

Format	Model	FC	NR	SO	CFS	UG
Arrow	LLaVA-NeXT	10.26	1.27	0.00	<u>88.39</u>	0.08
	LLaVA-RLHF	4.56	2.23	0.00	<u>93.05</u>	0.15
	LLaVA-CoT	79.93	0.92	0.08	15.84	3.23
	CogVLM2	33.63	0.63	0.48	<u>46.94</u>	18.31
	MiniGPT-4	0.58	3.75	0.00	<u>95.67</u>	0.00
	Qwen3-VL	97.18	2.82	0.00	0.00	0.00
Semi.	LLaVA-NeXT	12.51	0.93	0.00	14.01	<u>72.55</u>
	LLaVA-RLHF	14.21	1.47	0.00	29.89	<u>54.44</u>
	LLaVA-CoT	85.09	10.89	0.40	0.78	2.89
	CogVLM2	13.04	0.68	7.80	<u>75.45</u>	3.03
	MiniGPT-4	0.83	1.63	0.00	<u>96.55</u>	0.98
	Qwen3-VL	97.77	2.20	0.00	0.00	0.03
Num.	LLaVA-NeXT	98.25	1.72	0.00	0.02	0.02
	LLaVA-RLHF	87.81	1.87	0.03	0.02	10.28
	LLaVA-CoT	97.25	0.88	0.00	0.00	1.87
	CogVLM2	84.44	0.83	9.06	0.00	5.66
	MiniGPT-4	97.68	2.08	0.00	0.13	0.10
	Qwen3-VL	97.77	2.23	0.00	0.00	0.00
Adj. L.	LLaVA-NeXT	2.67	1.02	0.00	25.85	<u>70.46</u>
	LLaVA-RLHF	0.00	1.72	0.00	<u>49.44</u>	<u>48.85</u>
	LLaVA-CoT	71.60	1.00	0.92	22.76	3.73
	CogVLM2	13.79	3.62	12.49	<u>68.96</u>	1.15
	MiniGPT-4	48.39	1.63	5.31	30.00	14.66
	Qwen3-VL	<u>44.24</u>	2.53	0.00	<u>53.02</u>	0.20

format production failure is the primary source of low structural scores. Score-compliance pairing captures this pattern. Reporting scores without compliance would miss it. Whether this confound is equally severe in stronger instruction-following models is an open question, but within the tested capability range it is substantial.

4.3 Layer 3: Scaffolding Robustness

The compliance data in Table 2 provides a measure of scaffolding robustness, captured by the compliance drop from C1 to C3 for each format. Average drops across models are Arrow -24.9, Semicolon -46.5, Adjacency -29.2, and Numbered $+4.2$ (compliance *improves*). Numbered format is scaffolding-invariant across all architectures. All other formats are scaffolding-fragile for most models.

Architectural findings. Qwen3-VL achieves near-perfect scaffolding robustness for all linearized formats (compliance drops of 1.4–2.0 points from C1 to C3) but not for the adjacency list (97.9% $\rightarrow$ 44.2%, a drop of 53.7 points). The structural complexity boundary is simultaneously a scaffolding boundary. LLaVA-RLHF shows format-specific suppression. At C1, 99.97% of adjacency list outputs and 91.29% of arrow outputs are Complete Non-Responses, but removing examples (C2) recovers arrow performance to 65.0% compliance. CogVLM2 shows the widest gap. Text-Only scores reach 8.83 at L1 but structural compliance is below 50% for 9 of 12 format-scaffolding conditions.

Format-dependent failure mechanisms. Table 3 presents failure mode distributions at minimal scaffolding, classified through rule-based parsing (§3.5). The same model fails through qualitatively different mechanisms depending on format. LLaVA-NeXT at C3 yields 88.39% Cross-Format Substitution for arrows but 72.55% Ungrounded Structure for semicolons, while maintaining 98.25% compliance for numbered format. For arrows, the failure is format selection. The model defaults to a different notation. For semicolons, the model generates the correct notation but fills it with abstract placeholders. For numbered sequences, failure is near-zero. Three different failure mechanisms for the same model under the same scaffolding, differing only in the requested output format.

Numbered format is universally robust. At C3, all six models achieve >84% compliance for numbered sequences, including MiniGPT-4 at 97.68% (vs. 0.58% for arrows) and LLaVA-RLHF at 87.81% (vs. 4.56% for arrows). Cross-Format Substitution is absent for numbered (<0.15% for all models). Among the four formats, numbered sequences provide the least confounded assessment of causal competence. They are format-robust and scaffolding-robust across all architectures. Human evaluation on a 10-image subset confirms these patterns, with numbered scaffolding robustness and the linearized-to-adjacency complexity boundary both replicated (Appendix A.7).

Table 4: Fine-tuning results: baseline (B), single-format (SF), and multi-format multi-scaffolding (MF). L3 structural scores averaged across judges; subscripts show format compliance (%). SF trains on semicolon at partial scaffolding; MF trains on all formats and scaffolding levels. Bold = best per row.

(a) LLaVA-CoT

	B	SF	MF
Format robustness (partial scaffolding)			
Arrow	$7.15_{99.3}$	$7.41_{1.9}$	$\mathbf{7.71}_{99.9}$
Semi.	$3.06_{67.9}$	$\mathbf{7.95}_{99.9}$	$\mathbf{7.95}_{99.9}$
Num.	$7.45_{99.3}$	$7.47_{0.2}$	$\mathbf{8.19}_{99.9}$
Adj. L.	$5.73_{87.9}$	$5.47_{32.5}$	$\mathbf{8.03}_{99.7}$
Scaffolding robustness (semicolon)			
C1	$6.76_{98.5}$	7.83_{100}	$\mathbf{7.95}_{99.8}$
C2	$3.06_{67.9}$	$\mathbf{7.95}_{99.9}$	$\mathbf{7.95}_{99.9}$
C3	$5.08_{85.1}$	$7.93_{99.8}$	$\mathbf{7.97}_{99.9}$

(b) MiniGPT-4

	B	SF	MF
Format robustness (partial scaffolding)			
Arrow	$4.87_{13.1}$	$\mathbf{6.35}_{86.1}$	$5.74_{73.3}$
Semi.	$4.71_{36.1}$	$\mathbf{6.88}_{96.6}$	$6.51_{87.0}$
Num.	$6.06_{92.8}$	$6.82_{95.2}$	$\mathbf{7.33}_{99.1}$
Adj. L.	$2.98_{62.2}$	$1.83_{44.3}$	$\mathbf{2.58}_{54.3}$
Scaffolding robustness (semicolon)			
C1	$5.22_{66.5}$	$5.17_{61.1}$	$\mathbf{6.92}_{95.3}$
C2	$4.71_{36.1}$	$\mathbf{6.88}_{96.6}$	$6.51_{87.0}$
C3	$4.78_{0.8}$	$2.35_{14.9}$	$\mathbf{6.40}_{86.9}$

The adjacency list exposes the structural complexity boundary. Even Qwen3-VL, which maintains >97% compliance for all three linearized formats at C3, drops to 44.24% for adjacency lists, with 53.02% Cross-Format Substitution. LLaVA-NeXT shows the most extreme scaffolding fragility. Adjacency drops from 93.99% at C1 to 2.67% at C3, a 91-point gap, while numbered drops by only 1.4 points over the same range.

Opposite scaffolding polarities. SSI decomposition shows scaffolding effects are not uniform. For arrows at L3, LLaVA-RLHF shows $\text{SSI}_{\text{ex}} = -5.24$ (removing examples improves performance, resolving suppression), while LLaVA-NeXT shows $\text{SSI}_{\text{ex}} = +0.12$ (removing examples hurts). For semicolons, LLaVA-CoT shows a non-monotonic pattern. Compliance drops from 98.45% (C1) to 67.90% (C2) before recovering to 85.09% (C3). For the adjacency list, even Qwen3-VL becomes strongly scaffolding-dependent ($\text{SSI}_{\text{full}} = +2.83$ at L3). See Appendix A.9 for complete SSI values.

Format-scaffolding combinations where a model's structural quality collapses constitute the model's **blind spots**. These blind spots motivate the targeted interventions in §4.4.

4.4 Layer 4: Targeted Intervention

We select LLaVA-CoT and MiniGPT-4 for fine-tuning because they offer maximum architectural contrast (chain-of-thought supervision vs. standard projection) on a shared blind spot (semicolon C2, where both collapse for different reasons). The remaining four models are less informative for this intervention. Qwen3-VL is near ceiling for linearized formats, LLaVA-NeXT overlaps with LLaVA-CoT, LLaVA-RLHF's suppression complicates interpretation, and CogVLM2's broad weakness makes isolating format-specific effects difficult. Table 4 reports L3 structural quality paired with compliance across baseline (B), single-format single-scaffolding (SF), and multi-format multi-scaffolding (MF) training.

Stage 1: Single-Format, Single-Scaffolding. Single-format training creates a dissociation between the two robustness dimensions. LLaVA-CoT becomes scaffolding-robust (semicolon compliance 99.83–99.95% across all scaffolding levels; L3: 3.06 → 7.95 at C2) but format-fragile (arrow compliance: 99.28% → 1.87%, dominated by Cross-Format Substitution). Score-compliance pairing exposes what scores alone cannot. Arrow L3 quality is maintained at 7.41 despite 1.87% compliance, because the model generates high-quality causal content in semicolon notation when arrows are requested. The causal competence transferred, but the format channel broke. MiniGPT-4 shows

the opposite. Format robustness improves for linearized formats (arrow: 13.11% $\rightarrow$ 86.06%) but scaffolding robustness collapses (C3: 0.83% $\rightarrow$ 14.88%, 84.69% Structure Omission). L1 text quality improves by $\sim$+0.6 across all conditions for MiniGPT-4, which points to a broader reasoning benefit from structural training. Detailed analysis is in Appendix A.10.

Stage 2: Multi-Format, Multi-Scaffolding. For LLaVA-CoT, format robustness is fully restored. Compliance exceeds 99.67% across all four formats, with quality improvements for all non-semicolon formats (adjacency list L3: 5.47 $\rightarrow$ 8.03; numbered: 7.47 $\rightarrow$ 8.19). Scaffolding robustness is preserved. The adjacency list result is informative. At 8.03 with 99.67% compliance, LLaVA-CoT after MF training exceeds Qwen3-VL's baseline adjacency performance (6.41 at 82.91%). Baseline evaluation ranked LLaVA-CoT below Qwen3-VL on adjacency lists. This ranking reverses after training. For these two models, baseline rankings on difficult formats reflected training exposure, not architectural capacity.

For MiniGPT-4, scaffolding robustness is substantially achieved (C3 compliance: 14.88% $\rightarrow$ 86.89%; L3: 2.35 $\rightarrow$ 6.40). Format capacity, however, remains bounded. Arrow compliance drops from 86.06% (SF) to 73.33% (MF), and adjacency list compliance improves only modestly (44.29% $\rightarrow$ 54.34%). Complete failure mode distributions are in Appendix A.10.

The architectural asymmetry. For LLaVA-CoT, both fragilities were data bottlenecks, fully resolvable through training diversity. For MiniGPT-4, scaffolding fragility was largely a training signal limitation, but format capacity shows evidence of architectural constraints. The adjacency list is the clearest indicator of this boundary.

5 Discussion

The four-layer framework indicates that the structural expression of VLM causal reasoning is both format-fragile and scaffolding-dependent, and that these effects are separable from causal content quality. Models with fluent causal text scores above 8.0 can fall below 1% format compliance under a different structural notation. For five of six tested models, structural quality tracks textual quality when compliance is high. CogVLM2 is the exception. Its structural quality (4.46) remains below its textual quality (5.87) within the same response despite 84% compliance (CDD = 0.24), whereas the other five models have CDD below 0.07. Across the tested models, low structural scores are driven primarily by format production failure, not poor causal content.

The behavioral patterns suggest possible sources of these fragilities. The dominance of Cross-Format Substitution for arrow notation is consistent with models reverting to formats more common in pretraining data. MiniGPT-4's improvement in text quality (+0.63) after structural training suggests sharing between structural and textual representations. In LLaVA-CoT, both fragilities respond to greater training diversity, whereas in MiniGPT-4, scaffolding sensitivity improves with training but format capacity remains limited. After multi-format, multi-scaffolding training, LLaVA-CoT reaches 8.03 on adjacency lists, compared with Qwen3-VL's baseline of 6.41. For this architecture, robustness can improve substantially when the training distribution covers the forms encountered at evaluation. Whether this pattern extends to models with weaker instruction following or different architectures remains an open question.

The results suggest three practical considerations for evaluation design. Although derived from causal reasoning experiments, they apply whenever compliance varies across output conditions. First, task scores should be reported with format compliance, since a low score may reflect failure to generate the requested format, not task failure. Second, conclusions about capability should not rest on a single structural format when performance can vary sharply across representations. Third, varying scaffolding helps distinguish self-sustained performance from prompt-dependent performance.

Our framework is a behavioral diagnostic and does not identify internal mechanisms behind these effects. It does, however, provide hypotheses for future mechanistic studies. The six models span a 2023–2025 capability range, hence the magnitude of the reported compliance gaps should not be assumed to hold for newer systems. Stronger instruction following may reduce sensitivity to some linearized formats, although Qwen3-VL still shows a 53-point compliance gap for adjacency lists despite near-zero linearized sensitivity. The score-compliance methodology is not tied to a particular model scale, but the empirical gaps reported here are specific to the models and conditions tested. Full limitations are provided in Appendix A.11.

AI USE STATEMENT

Generative AI tools were used in this work in three capacities.

As part of the research methodology. GPT-4o was used to generate causal questions and ground-truth structures for the evaluation and training datasets (Section 3). Claude 4.5 Sonnet and Gemini 2.5 Flash were used as automated evaluation judges (Section 3.5). These uses are part of the experimental methodology and are described and cited in the corresponding sections.

For manuscript preparation. Generative AI tools assisted with drafting, editing, and iterative refinement of the manuscript, including articulation, grammar, readability, and organization.

For literature analysis. AI tools assisted with summarizing potentially relevant literature and checking whether cited claims were consistent with the original publications.

Author contributions and oversight. The research questions, conceptual framing, benchmark design, experimental pipeline, evaluation infrastructure, implementation, fine-tuning experiments, statistical analyses, and scientific conclusions were developed by the authors. AI-generated suggestions concerning methodology, experiments, or interpretation were considered critically and checked against experimental results and primary sources before any decision was made. Apart from the explicitly described use of GPT-4o for dataset construction, generative AI was not used to derive experimental results, reported metrics, scientific figures, or code for the experimental pipeline.

All AI-assisted material was reviewed by the authors, who take responsibility for the final manuscript, including its claims, analyses, citations, and artifacts.

ETHICS STATEMENT

This paper evaluates publicly available VLMs using images from MS-COCO, a widely used research dataset. No human-subjects research was conducted beyond expert annotation performed by the authors. The causal questions concern physical and environmental scenarios and do not target individuals. The evaluation framework is designed to improve the reliability of VLM causal reasoning assessment. We do not anticipate substantial societal risks arising from the study.

REPRODUCIBILITY STATEMENT

The evaluation and training datasets are constructed from MS-COCO as described in Section 3. All six evaluated VLMs are publicly available open-source models. Fine-tuning hyperparameters are provided in Appendix A.5, with the same settings used across training stages for each model. The rule-based regex procedure used to assign failure modes is described in Section 3.5. Per-judge scores for all experimental conditions are reported in Appendix A.6, and the human validation protocol and results are provided in Appendix A.7. Code and data will be released upon acceptance.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 9525–9536, Red Hook, NY, USA, 2018. Curran Associates Inc.

Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. Faithfulness vs. plausibility: On the (un)reliability of explanations from large language models, 2024. URL https://arxiv.org/abs/2402.04614.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

David Alvarez-Melis and Tommi S. Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 7786–7795, Red Hook, NY, USA, 2018. Curran Associates Inc.

Anthropic. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, September 2025. Accessed: 2026-01-24.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. A closer look at bias and chain-of-thought faithfulness of large (vision) language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13406–13439, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.723. URL https://aclanthology.org/2025.findings-emnlp.723/.

Mohamad Ballout, Okajevo Wilfred, Seyedalireza Yaghoubi, Nohayr Muhammad Abdelmoneim, Julius Mayer, and Elia Bruni. Can you SPLICE it together? a human curated benchmark for probing visual reasoning in VLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11288–11309, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. URL https://aclanthology.org/2025.findings-emnlp.604/.

Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013. doi: 10.1073/pnas.1306572110. URL https://www.pnas.org/doi/abs/10.1073/pnas.1306572110.

Oana-Maria Camburu, Brendan Shillingford, Pasquale Minervini, Thomas Lukasiewicz, and Phil Blunsom. Make up your mind! adversarial generation of inconsistent natural language explanations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4157–4165, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.382. URL https://aclanthology.org/2020.acl-main.382/.

Meiqi Chen, Bo Peng, Yan Zhang, and Chaochao Lu. CELLO: Causal evaluation of large vision-language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22353–22374, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1247. URL https://aclanthology.org/2024.emnlp-main.1247/.

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020.

Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11 (2):127–138, 2010. doi: 10.1038/nrn2787. URL https://doi.org/10.1038/nrn2787.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025. URL https://arxiv.org/abs/2411.15594.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024. doi: 10.1109/CVPR52733.2024.01363.

Ian Hartsock and Ghulam Rasool. Vision–language models for medical report generation and visual question answering: A review. *Frontiers in Artificial Intelligence*, 7:1430984, 2024. doi: 10.3389/frai.2024.1430984.

Xanh Ho, Jiahao Huang, Florian Boudin, and Akiko Aizawa. Llm-as-a-judge: Reassessing the performance of llms in extractive qa, 2025. URL https://arxiv.org/abs/2504.11972.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, Lei Zhao, Zhuoyi Yang, Xiaotao Gu, Xiaohan Zhang, Guanyu Feng, Da Yin, Zihan Wang, Ji Qi, Xixuan Song, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Yuxiao Dong, and Jie Tang. Cogvlm2: Visual language models for image and video understanding, 2024. URL https://arxiv.org/abs/2408.16500.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Aneesh Komanduri, Karuna Bhaila, and Xintao Wu. CausalVLBench: Benchmarking visual causal reasoning in large vision-language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 30660–30680, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1561. URL https://aclanthology.org/2025.emnlp-main.1561/.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73, May 2017. ISSN 0920-5691. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning, 2023. URL https://arxiv.org/abs/2307.13702.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, June 2023a. URL http://arxiv.org/abs/2301.12597. arXiv:2301.12597 [cs].

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20.

Zhiyuan Li, Heng Wang, Dongnan Liu, Chaoyi Zhang, Ao Ma, Jieting Long, and Weidong Cai. Multimodal causal reasoning benchmark: Challenging vision large language models to discern causal links across modalities, 2025. URL https://arxiv.org/abs/2408.08105.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Do Xuan Long, Ngoc-Hai Nguyen, Tiviatis Sim, Hieu Dao, Shafiq Joty, Kenji Kawaguchi, Nancy F. Chen, and Min-Yen Kan. LLMs are biased towards output formats! systematically evaluating and mitigating output format bias of LLMs. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 299–330, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.15. URL https://aclanthology.org/2025.naacl-long.15/.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723, June 2024. doi: 10.1162/coli_a_00511. URL https://aclanthology.org/2024.cl-2.6/.

Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 295–337, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.19. URL https://aclanthology.org/2024.findings-acl.19/.

OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. URL https://openai.com/index/hello-gpt-4o/. Accessed: 2025-12-02.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009. ISBN 9780521895606. doi: 10.1017/CBO9780511803161. URL https://doi.org/10.1017/CBO9780511803161.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning, 2021. URL https://arxiv.org/abs/2102.11107.

Luca M. Schulze Buschoff, Elif Akata, Matthias Bethge, and Eric Schulz. Visual cognition in multimodal large language models. *Nature Machine Intelligence*, 7:96–106, 2025. doi: 10.1038/s42256-024-00963-y. URL https://doi.org/10.1038/s42256-024-00963-y.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, 2024. URL https://arxiv.org/abs/2310.11324.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2001.

Xin Su, Man Luo, Kris W Pan, Tien Pei Chou, Vasudev Lal, and Phillip Howard. Sk-vqa: Synthetic knowledge generation at scale for training context-augmented multimodal llms, 2025. URL https://arxiv.org/abs/2406.19593.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. URL https://arxiv.org/abs/2309.14525.

Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 1218–1236, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.91. URL https://aclanthology.org/2024.emnlp-industry.91/.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Martin Tutek, Fateme Hashemi Chaleshtori, Ana Marasovic, and Yonatan Belinkov. Measuring chain of thought faithfulness by unlearning reasoning steps. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9935–9960, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.504. URL https://aclanthology.org/2025.emnlp-main.504/.

Zeqing Wang, Shiyuan Zhang, Chengpei Tang, and Keze Wang. Timecausality: Evaluating the causal ability in time dimension for vision language models, 2025. URL https://arxiv.org/abs/2505.15435.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL https://aclanthology.org/D19-1002/.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024. URL https://arxiv.org/abs/2411.10440.

Jialin Yang, Dongfu Jiang, Lipeng He, Sherman Siu, Yuxuan Zhang, Disen Liao, Zhuofeng Li, Huaye Zeng, Yiming Jia, Haozhe Wang, Benjamin Schneider, Chi Ruan, Wentao Ma, Zhiheng Lyu, Yifei Wang, Yi Lu, Quy Duc Do, Ziyan Jiang, Ping Nie, and Wenhu Chen. Structeval: Benchmarking llms' capabilities to generate structural outputs, 2026. URL https://arxiv.org/abs/2505.20139.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal, 2023. URL https://arxiv.org/abs/2308.13067.

Jusheng Zhang, Kaitong Cai, Xiaoyang Guo, Sidi Liu, Qinhan Lv, Ruiqi Chen, Jing Yang, Yijia Fan, Xiaofei Sun, Jian Wang, Ziliang Chen, Liang Lin, and Keze Wang. Mm-cot:a benchmark for probing visual chain-of-thought reasoning in multimodal models, 2025. URL https://arxiv.org/abs/2512.08228.

Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. What if the tv was off? examining counterfactual reasoning abilities of multi-modal language models, 2024. URL https://arxiv.org/abs/2310.06627.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C. Knoll. Vision language models in autonomous driving: A survey and outlook, 2024a. URL https://arxiv.org/abs/2310.14414.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models, 2024b. URL https://arxiv.org/abs/2310.00754.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Appendix

A.1 Extended Related Work

A.1.1 Faithfulness vs. Plausibility in Model Explanations

The distinction between faithful and plausible explanations has emerged as a central challenge in explainable AI. Faithful explanations accurately represent a model's internal reasoning. Plausible explanations seem convincing to humans without necessarily reflecting actual model processing (Jacovi & Goldberg, 2020; Lyu et al., 2024). Agarwal et al. (2024) argue demand for user-friendly interfaces often increases plausibility at the cost of faithfulness, and models can generate highly plausible explanations with no correspondence to actual reasoning pathways. Wiegreffe & Pinter (2019) show attention mechanisms can be plausible without being faithful, and Camburu et al. (2020) find models trained to generate explanations often default to post-hoc rationalizations. Madsen et al. (2024) show the faithfulness of LLM self-explanations varies substantially by task, model, and explanation type.

Methods for evaluating faithfulness include perturbation-based approaches (Alvarez-Melis & Jaakkola, 2018; Hooker et al., 2019), which test whether removing features identified as important degrades performance, and consistency-based approaches (Adebayo et al., 2018; DeYoung et al., 2020), which assess whether explanations remain stable under perturbations. These strategies are designed for gradient-based or attention-based explanations in NLP and do not naturally extend to structured outputs such as causal graphs generated by VLMs.

For chain-of-thought reasoning specifically, Lanham et al. (2023) propose tests including early answering, mistake insertion, and paraphrasing to assess whether CoT traces are faithful. Turpin et al. (2023) show models can generate correct CoT reasoning while being influenced by biasing features absent from the CoT. Fluent reasoning traces do not guarantee faithfulness. Tutek et al. (2025) introduce a parametric approach (FUR) for erasing information from reasoning steps and measuring the effect on predictions. Unlearning steps frequently changes model outputs. Balasubramanian et al. (2025) present the first study of CoT faithfulness in vision-language models. Their findings show image-based biases are rarely articulated in reasoning traces, even by models specialized for reasoning. They also identify "inconsistent reasoning" where models reason correctly before abruptly changing answers.

Our work extends this literature in a different direction. Instead of testing whether existing CoT traces are faithful to internal processing, we test whether models can express causal reasoning consistently across multiple structural formats and scaffolding conditions. Format-specific fragility and scaffolding dependence provide complementary behavioral diagnostics. When a model generates fluent causal text but cannot maintain structural output quality across formats or without prompt support, the capability does not hold.

A.1.2 Output Format Sensitivity in Language Models

Recent research has established LLM and VLM performance is highly sensitive to output format specifications. The question is whether models possess genuine task understanding or rely on format-specific pattern matching.

Sclar et al. (2024) show performance can vary by up to 76 accuracy points across semantically equivalent prompt formats in few-shot settings, and this sensitivity persists despite instruction tuning and model scaling. Tam et al. (2024) investigate the impact of structured format restrictions (JSON, XML, YAML) on LLM reasoning and find degradation under format constraints, with stricter formats causing greater performance loss. They observe format restrictions can create competing goals for the model between reasoning and format adherence. Long et al. (2025) systematically evaluate output format bias across wrapping strategies, list formats, and mapping structures. LLMs exhibit systematic and model-specific biases toward particular formats. The authors introduce metrics to disentangle format compliance from task accuracy. Yang et al. (2026) benchmark structured output generation across 18 formats. Even GPT-4o achieves only 76% accuracy, with generation tasks consistently harder than conversion tasks and visual rendering harder than text-only structures.

These findings are relevant to our work. If a VLM fails to generate a causal chain in arrow notation, prior causal reasoning evaluations would conclude the model lacks causal capability. Our multi-format evaluation shows this conclusion may be premature. The same model may generate high-quality causal structures in a different format. The format sensitivity established for LLMs in general extends to VLM causal reasoning, and our framework provides diagnostic tools to measure format robustness and separate format production difficulty from causal competence.

A.1.3 VLM Causal Reasoning Benchmarks

Benchmarks for evaluating causal reasoning in VLMs have proliferated, organized around Pearl's Ladder of Causation (Pearl, 2009; Pearl & Mackenzie, 2018).

CELLO (Chen et al., 2024) evaluates VLMs across four causal levels using multiple-choice questions with predefined causal graphs from Visual Genome scene graphs (Krishna et al., 2017). While broad in scope, the multiple-choice format allows models to select from options without generating causal structure, and performance reflects selection capability, not generation capability.

MM-CoT (Zhang et al., 2025) probes visual chain-of-thought reasoning with 5,616 image-based instances requiring models to select valid causal chains from distractors. The benchmark finds models struggle with chain selection, and semantic redundancy (repeating descriptions without advancing the causal chain) is the most common error type. Our work extends beyond selection. We require models to generate causal chains, and we show generation performance depends on the format of the requested output.

CausalVLBench (Komanduri et al., 2025) evaluates causal structure inference, intervention target prediction, and counterfactual prediction using synthetic image pairs with known causal graphs. The benchmark provides important controlled evaluation but uses synthetic stimuli and multiple-choice or binary response formats. Our evaluation uses natural scene images and requires open-ended structural generation.

Zhang et al. (2024) examine counterfactual reasoning through object state manipulation and show VLMs struggle with simple counterfactual changes despite fluent responses. Li et al. (2025) evaluate cross-modal causal reasoning with 12,000 synthetic image pairs and propose visual chain-of-thought prompting. Wang et al. (2025) test temporal causal reasoning and find performance degrades when models infer temporal order without explicit prompting, a finding consistent with our scaffolding sensitivity results. Ballout et al. (2025) find VLMs struggle with event-based reasoning in instructional videos, relying on language priors over visual understanding.

A common thread across these benchmarks is each uses a single output modality (free text, multiple-choice, or one structural format) and characterizes weaknesses without diagnosing their source. Our work adds both the format and scaffolding dimensions to causal reasoning evaluation, and uses targeted fine-tuning to determine whether identified fragilities reflect training signal limitations or architectural constraints.

A.1.4 CAUSAL REASONING IN AI

Causal reasoning is a core aspect of intelligence, requiring understanding of cause-and-effect relationships for planning, intervention, and counterfactual reasoning (Pearl & Mackenzie, 2018; Pearl, 2009). Causal inference has been studied extensively in structured domains (Spirtes et al., 2001), and recent deep learning approaches integrate causal principles in representation learning (Schölkopf et al., 2021) and counterfactual reasoning. Cognitive science research suggests genuine causal understanding relies on internal generative models representing latent causes (Friston, 2010; Battaglia et al., 2013). If models truly understand causality, they should be able to abstract relationships into structural form, beyond generating convincing causal language. Our work uses structural generation as a complementary diagnostic for causal competence, without treating it as a prerequisite for causal understanding. The finding that structural capability is format-fragile and scaffolding-dependent indicates whatever causal competence these models possess, its expression through structural output does not hold uniformly.

A.1.5 VLM CHALLENGES AND HALLUCINATION

VLMs show impressive performance in captioning, VQA, and instruction following (Liu et al., 2023; Alayrac et al., 2022; Li et al., 2023a), but critical challenges persist. Models generate plausible descriptions while ignoring visual evidence (Guan et al., 2024; Tong et al., 2024), generate hallucinated content for straightforward visual questions (Li et al., 2023b), and rely on language priors over visual grounding (Zhou et al., 2024b; Sun et al., 2023). VLM causal judgments sometimes correlate negatively with human responses (Schulze Buschoff et al., 2025). These findings indicate fluent causal language generation can coexist with poor structural understanding. Our work shows this gap is not a fixed model property but is format-contingent and scaffolding-dependent. The same model can score well under one format and poorly under another, with format compliance varying from near-ceiling to near-zero depending on the requested notation and the level of prompt support.

A.2 DIAGNOSTIC METRICS: EXTENDED MOTIVATION

Standard evaluation metrics report overall accuracy or quality scores but do not capture how model behavior varies across structural formats, reasoning depths, and scaffolding conditions. We introduce diagnostic metrics designed to quantify specific dimensions of evaluation reliability. These metrics are not intended to replace quality scores from LLM judges but to provide complementary information about conditions under which model performance changes.

Abstraction Gap (AG). Requiring a structural representation may influence the quality of the accompanying textual explanation. AG measures the difference between the text score in the Text-Only Probe and the text score in the Structure-Text Probe, where the model generates a causal structure before generating the explanation. This metric captures whether the requirement to express reasoning in a structured form affects the model's ability to generate coherent causal explanations. Prior work on output format constraints supports the view reasoning performance can depend on the requested representation (Tam et al., 2024; Long et al., 2025).

Causal Depth Disparity (CDD). Even when models generate both a textual explanation and a corresponding causal structure, the two may not reflect the same level of reasoning. CDD measures the normalized difference between the textual and structural scores within a single response ($\text{CDD} = (S_{\text{ST}}^{\text{text}} - S_{\text{ST}}^{\text{struct}})/S_{\text{ST}}^{\text{text}}$). Because our judges score causal content quality regardless of format compliance, CDD is interpretable only when compliance is high. At low compliance, the structural score reflects the quality of whatever format the model actually generated (typically Cross-Format Substitution), not the quality of the requested format. In practice, CDD approaches zero at high-compliance conditions for five of six models. The measured gap between plausible text and structural output is driven by format production failure, not by a divergence in reasoning quality. CDD

remains useful for identifying models where a quality gap persists even at accessible formats (e.g., CogVLM2). The metric relates to the broader distinction between plausible and faithful explanations (Jacovi & Goldberg, 2020; Lyu et al., 2024; Agarwal et al., 2024).

Format Variance (FV). The same causal content can be represented in multiple structural formats. If a model's causal reasoning is format-robust, performance should remain consistent across these equivalent representations. FV measures the variability of structural scores across the four formats. It captures sensitivity to notational representation changes. Prior work on prompt sensitivity and evaluation reliability motivates this metric (Sclar et al., 2024; Yang et al., 2026).

Scaffolding Sensitivity Index (SSI). Prompt scaffolding techniques such as worked examples and step-by-step instructions are commonly used to improve reasoning performance. However, strong improvements under heavily guided prompts may reflect reliance on prompt structure over scaffolding-independent capability. SSI measures how performance changes across different levels of prompt guidance, and its decomposition into SSI_{ex} (example removal) and $\text{SSI}_{\text{instr}}$ (instruction removal) isolates the contribution of each scaffolding component. Large positive SSI indicates scaffolding-dependent capability, externally conditioned and not self-sufficient (Sclar et al., 2024; Lanham et al., 2023).

A.3 Prompt Templates

We present the complete prompt templates for all 12 Structure-Text Probe conditions (4 formats $\times$ 3 scaffolding levels). Each template uses $\{\}$ as a placeholder for the causal question. All prompts include the image as visual input. The Text-Only Probe uses the same question with a simple instruction to answer in free text, with no structural output requested.

The scaffolding reduction follows a consistent pattern across all four formats. C1 (full) includes Pearl's causal framework reference, detailed construction instructions, format specification, and three worked examples. C2 (partial) removes the worked examples. C3 (minimal) removes both the Pearl framework reference and the detailed construction instructions, retaining only a brief task description with format syntax cues.

We present the arrow format in full across all three scaffolding levels to illustrate this reduction. For semicolon and numbered formats, we present C1 only, as C2 and C3 follow the identical reduction pattern with format-specific notation substituted. For the adjacency list, we present both C1 and C3 to allow verification of the prompt asymmetry discussed in §A.11.

A.3.1 Arrow Format

C1 (Full Scaffolding).

Arrow C1 Prompt

You are a vision-language model (VLM) tasked with answering a provided question about an image in a way that demonstrates causal reasoning, based on Judea Pearl's Ladder of Causation: Level 1 (Association) identifies observable scene patterns; Level 2 (Intervention) predicts outcomes of a specific change, considering temporal/environmental factors; Level 3 (Counterfactual) reasons about alternate scenarios if a past condition differed, requiring multi-step causal logic.
You are required to construct a lightweight causal chain to represent the chain of cause-and-effect relationships relevant to the answer, using the number of nodes and arrows needed to accurately capture the causal logic (e.g., 2 nodes like A $\rightarrow$ B, 3 nodes like A $\rightarrow$ B $\rightarrow$ C, or more if appropriate), tailored to the complexity of the reasoning. You must provide the causal chain first, followed by a concise text response that explains the reasoning based on the chain.
Instructions:

- Base the response on the image and question: Ensure the answer and chain are grounded in the image's content and relevant to the question.

Arrow C1 Prompt (continued)

- Constructing the causal chain: Follow these steps: (1) Identify the key cause-and-effect relationships implied by the question. (2) Format these relationships as a chain using '→' to connect nodes (e.g., Rain → Wet Ground → Darkened Scene). (3) Use the fewest nodes needed for clarity, but include additional nodes if the causal chain requires multiple steps.
- Provide a concise answer: Write 1–2 sentences that explain the causal reasoning, referencing the chain's steps.
- Avoid overly simplistic chains: Ensure the response and chain capture specific, relevant causal relationships.
- Format the output exactly as follows:

Causal Chain: [Your causal chain, e.g., A → B or A → B → C → D]
Answer: [Your answer, explaining the reasoning based on the chain]
Example:
Question: How would the scene change if a sudden rainstorm began in a sunny park?
Causal Chain: Rainstorm → Wet Ground → Darkened Atmosphere
Answer: The rainstorm wets the ground, creating a darker, moodier atmosphere as rain falls.
Question: What would the park look like if it had been photographed at night instead of midday?
Causal Chain: Night → Reduced Natural Light
Answer: At night, the lack of natural light darkens the park, with only artificial lights visible.
Question: How would the scene differ if a month-long drought had preceded this day?
Causal Chain: Drought → Dry Soil → Wilted Vegetation → Sparse Leaves
Answer: The drought dries the soil, leading to wilted, brown grass and sparse tree leaves.
Provided Information:
Question: {}
Task: Answer the provided question based on the image, and always provide a lightweight causal chain to represent the causal reasoning. Follow the instructions above and output in the appropriate format:
Causal Chain: [Your causal chain]
Answer: [Your answer]

C2 (Partial Scaffolding). Identical to C1 with the three worked examples removed. The Pearl framework reference, construction instructions, and format specification are retained.

Arrow C2 Prompt

You are a vision-language model (VLM) tasked with answering a provided question about an image in a way that demonstrates causal reasoning, based on Judea Pearl's Ladder of Causation: Level 1 (Association) identifies observable scene patterns; Level 2 (Intervention) predicts outcomes of a specific change, considering temporal/environmental factors; Level 3 (Counterfactual) reasons about alternate scenarios if a past condition differed, requiring multi-step causal logic.
You are required to construct a lightweight causal chain to represent the chain of cause-and-effect relationships relevant to the answer, using the number of nodes and arrows needed to accurately capture the causal logic (e.g., 2 nodes like A → B, 3 nodes like A → B → C, or more if appropriate), tailored to the complexity of the reasoning. You must provide the causal chain first, followed by a concise text response that explains the reasoning based on the chain.
Instructions: [Identical to C1]
Causal Chain: [Your causal chain, e.g., A → B or A → B → C → D]
Answer: [Your answer, explaining the reasoning based on the chain]
Provided Information:
Question: {}
Task: Answer the provided question based on the image, and always provide a lightweight causal chain to represent the causal reasoning. Follow the instructions above and output in the appropriate format:
Causal Chain: [Your causal chain]
Answer: [Your answer]

C3 (Minimal Scaffolding). Removes Pearl's framework reference and detailed construction instructions. Retains the task description, format syntax cues, and output labels.

Arrow C3 Prompt

You are a vision-language model (VLM) tasked with answering a provided question about an image in a way that demonstrates causal reasoning.
You are required to construct a lightweight causal chain to represent the chain of cause-and-effect relationships relevant to the answer, using the number of nodes and arrows needed to accurately capture the causal logic (e.g., 2 nodes like A $\rightarrow$ B, 3 nodes like A $\rightarrow$ B $\rightarrow$ C, or more if appropriate), tailored to the complexity of the reasoning. You must provide the causal chain first, followed by a concise text response that explains the reasoning based on the chain.
Provided Information:
Question: {}
Task: Answer the provided question based on the image, and always provide a lightweight causal chain to represent the causal reasoning. Follow the instructions above and output in the appropriate format:
Causal Chain: [Your causal chain]
Answer: [Your answer]

A.3.2 Semicolon Format (C1 Only)

C2 and C3 follow the same reduction pattern as arrow, with semicolon notation substituted throughout.

Semicolon C1 Prompt

[Pearl framework reference: identical to Arrow C1]
You are required to construct a simple semicolon-separated sequence to represent the chain of cause-and-effect relationships relevant to the answer, using the number of elements needed to accurately capture the causal logic (e.g., 2 elements like A; B, 3 elements like A; B; C, or more if appropriate), tailored to the complexity of the reasoning. You must provide the causal semicolon-separated sequence first, followed by a concise text response that explains the reasoning based on the sequence.
Instructions: [Same structure as Arrow C1, with:] (2) Format these relationships as a sequence using ';' to connect elements (e.g., Rain; Wet Ground; Darkened Scene).
Causal Sequence: [Your causal sequence, e.g., A; B or A; B; C; D]
Answer: [Your answer]
Examples:
Q: How would the scene change if a sudden rainstorm began in a sunny park?
Causal Sequence: Rainstorm; Wet Ground; Darkened Atmosphere
Answer: The rainstorm wets the ground, creating a darker, moodier atmosphere as rain falls.
[Two additional examples with semicolon notation, same content as Arrow C1 examples]
Question: {}
Task: Answer the provided question based on the image, and always provide a simple semicolon-separated sequence to represent the causal reasoning.
Causal Sequence: [Your causal sequence]
Answer: [Your answer]

A.3.3 Numbered Format (C1 Only)

C2 and C3 follow the same reduction pattern as arrow, with numbered notation substituted throughout.

Numbered C1 Prompt

[Pearl framework reference: identical to Arrow C1]
You are required to construct a simple numbered sequence to represent the chain of cause-and-effect relationships relevant to the answer, using the number of elements needed to accurately capture the causal logic (e.g., 2 elements like 1. A 2. B, 3 elements like 1. A 2. B 3. C, or more if appropriate), tailored to the complexity of the reasoning. You must provide the causal numbered sequence first, followed by a concise text response that explains the reasoning based on the sequence.

Numbered C1 Prompt (continued)

Instructions: [Same structure as Arrow C1, with:] (2) Format these relationships as a numbered sequence (e.g., 1. Rain 2. Wet Ground 3. Darkened Scene).
Causal Sequence: [Your causal sequence, e.g., 1. A 2. B or 1. A 2. B 3. C 4. D]
Answer: [Your answer]
Examples:
Q: How would the scene change if a sudden rainstorm began in a sunny park?
Causal Sequence: 1. Rainstorm 2. Wet Ground 3. Darkened Atmosphere
Answer: [Same as Arrow]
[Two additional examples with numbered notation]
Question: {}
Task: Answer the provided question based on the image, and always provide a simple numbered sequence to represent the causal reasoning.
Causal Sequence: [Your causal sequence]
Answer: [Your answer]

A.3.4 Adjacency List Format

The adjacency list format is qualitatively different from the three linearized formats. We present C1 and C3 to allow verification of the prompt asymmetry discussed in §A.11.

C1 (Full Scaffolding).

Adjacency List C1 Prompt

[Pearl framework reference: identical to Arrow C1]
You are required to represent the causal relationships as an adjacency list (parent-to-direct-child mapping). Each line should name one node followed by a colon and then its immediate causal consequence(s) (e.g., A: B if A causes B). Use one line per parent node. For nodes with no outgoing edges (terminal effects), write "(none)" (e.g., A: (none)). List the nodes in a logical topological order (causes before effects). This format explicitly captures the directed graph structure, not just a linear chain.
You must provide the adjacency list first, followed by a concise text response that explains the reasoning based on the graph.
Instructions:

- Base the response on the image and question.
- Constructing the adjacency list: (1) Identify the key causal relationships. (2) Break them into direct cause → immediate effect pairs. (3) For each node with a consequence, write: "Node: Consequence". (4) For final effects, write: "Final Node: (none)". (5) Include all relevant nodes (usually 2–5 lines). (6) Order lines so causes appear before effects (topological order).
- Provide a concise answer: Write 1–2 sentences explaining the causal reasoning.
- Avoid overly simplistic graphs.

Format the output exactly as follows:
Causal Graph:
NodeA: NodeB
NodeB: NodeC
NodeC: (none)
Answer: [Your concise explanation]
Examples:
Q: How would the scene change if a sudden rainstorm began in a sunny park?
Causal Graph:
Rainstorm: Wet Ground
Wet Ground: Darkened Atmosphere
Darkened Atmosphere: (none)
Answer: The rainstorm wets the ground, which darkens the overall atmosphere.
Q: What would the park look like if photographed at night instead of midday?
Causal Graph:
Nightfall: Reduced Natural Light
Reduced Natural Light: Dark Park Scene
Dark Park Scene: (none)
Answer: At night, the absence of natural light causes the park to become significantly darker.

Adjacency List C1 Prompt (continued)

Q: How would the scene differ if a month-long drought had preceded this day?
Causal Graph:
Drought: Dry Soil
Dry Soil: Wilted Vegetation
Wilted Vegetation: Sparse Brown Leaves
Sparse Brown Leaves: (none)
Answer: The prolonged drought would dry out the soil, leading to wilted vegetation and sparse, brown leaves.
Question: {}
Task: Answer the provided question based on the image, and always provide the causal relationships in adjacency list format (one parent-to-child mapping per line).
Causal Graph:
[Your adjacency list lines here]
Answer: [Your answer]

C3 (Minimal Scaffolding). This prompt retains more instructional content than the other three formats at C3, because the adjacency list notation requires specifying the parent-child mapping convention, the "(none)" terminal marker, and the topological ordering instruction.

Adjacency List C3 Prompt

You are a vision-language model (VLM) tasked with answering a provided question about an image in a way that demonstrates causal reasoning.
You are required to represent the causal relationships as an adjacency list (parent-to-direct-child mapping). Each line should name one node followed by a colon and then its immediate causal consequence(s) (e.g., A: B if A causes B). Use one line per parent node. For nodes with no outgoing edges (terminal effects), write "(none)" (e.g., A: (none)). List the nodes in a logical topological order (causes before effects). This format explicitly captures the directed graph structure, not just a linear chain.
You must provide the adjacency list first, followed by a concise text response that explains the reasoning based on the graph.
Provided Information:
Question: {}
Task: Answer the provided question based on the image, and always provide the causal relationships in adjacency list format (one parent-to-child mapping per line).
Causal Graph:
[Your adjacency list lines here]
Answer: [Your answer]

Scaffolding reduction summary. Table 5 summarizes the components present at each scaffolding level across formats.

A.4 Dataset Construction and Example

Figure 1 illustrates an evaluation instance. For each image, GPT-4o generates nine causal question-answer pairs across Pearl's three causal levels, associative (L1), interventional (L2), and counterfactual (L3), with three questions per level. Causal structures are generated in arrow notation. The remaining three formats are obtained via rule-based conversion.

A.5 Fine-Tuning Configuration

Training data. Both stages use causal questions generated by GPT-4o for 5,000 MS-COCO training images (no overlap with the 1,000 evaluation images). Each image has 9 question-answer pairs (3 per causal level), yielding 45,000 total training examples. In Stage 1, all L2 and L3 examples use semicolon-format structures at partial scaffolding. In Stage 2, the L2 and L3 questions (30,000 total) are distributed equally across all twelve format-scaffolding combinations (2,500 per combination), cycling systematically so that every two images cover all combinations. L1 questions provide text-

Table 5: Prompt components by scaffolding level. ✓ = present. The adjacency list retains more specification at C3 than linearized formats.

Component	C1	C2	C3
All four formats			
Task description	✓	✓	✓
Format syntax cues	✓	✓	✓
Output labels	✓	✓	✓
Removed at C2			
Worked examples (×3)	✓		
Removed at C3			
Pearl framework ref.	✓	✓	
Construction instructions	✓	✓	
Adjacency list only at C3			
Parent-child convention	✓	✓	✓
"(none)" terminal marker	✓	✓	✓
Topological ordering	✓	✓	✓



Level	Question	Causal Structure	Textual Answer
1	What is the color of the leftmost beach umbrella?	None	White and Orange.
2	If a strong wind began to blow from the direction of the mountains, what would happen to the beach umbrellas?	Strong wind → Force on umbrellas → Umbrellas topple or fly away	The umbrellas may topple or fly away.
3	Had this photo been taken during a holiday weekend instead of what appears to be a quiet day, how would the beach scene differ?	Holiday weekend → More visitors → Crowded beach → Fewer visible umbrellas	The beach would be more crowded with fewer visible umbrellas.

Figure 1: Example evaluation instance illustrating Pearl's three causal levels. Level 1 (association) requires no causal structure. Levels 2–3 (intervention, counterfactual) require structure-first generation of explicit causal chains. This example shows arrow notation; semicolon, numbered, and adjacency list formats are also tested to assess format robustness.

only responses in both stages. Ground-truth structures are generated in arrow notation by GPT-4o and converted to the other three formats via rule-based conversion.

Hyperparameters. All hyperparameters are identical across Stage 1 and Stage 2 for each model. Table 6 reports the configuration.

Table 6: Fine-tuning hyperparameters. Values are identical across Stage 1 and Stage 2.

Parameter	LLaVA-CoT	MiniGPT-4
Trainable components	Connector, ViT, LLM decoder	Projection layer only
LoRA rank	128	N/A
LoRA alpha	256	N/A
Learning rate	2e-4	2e-4
Weight decay	0.01	0.01
Epochs	1	1
Training steps	3,000	3,000
Max sequence length	2,048	2,048
Loss function	Cross-entropy	Cross-entropy

LLaVA-CoT uses LoRA for parameter-efficient adaptation of the LLM decoder and vision encoder, with rank 128 and alpha 256. MiniGPT-4 fine-tunes the linear projection layer with full gradient updates, following the original training protocol where the Vicuna LLM remains frozen (Zhu et al., 2023). Both models train for a single epoch (3,000 steps) with identical learning rate, weight decay, and sequence length, ensuring that differences in outcome between the two models and between the two stages are attributable to architecture and format diversity, not training duration or optimization configuration.

A.6 Baseline Detailed Results

Table 7 reports per-judge Text-Only Probe scores. Tables 8–11 report per-judge L2 and L3 scores (Text / Structure) and failure mode distributions for all six models across all scaffolding conditions. Scores are on a 0–10 scale. Failure modes are percentages: FC = Format-Compliant, NR = Complete Non-Response, SO = Structure Omission, CFS = Cross-Format Substitution, UG = Ungrounded Structure. The main paper (Tables 1 and 2) reports averaged scores; these tables provide the full per-judge breakdown.

Table 7: Text-Only Probe: per-judge scores across all causal levels.

	Claude 4.5 Sonnet			**Gemini 2.5 Flash**		
	L1	L2	L3	L1	L2	L3
LLaVA-NeXT	7.74	7.14	7.27	8.62	9.08	8.67
LLaVA-RLHF	4.02	6.36	6.61	4.03	8.11	8.21
LLaVA-CoT	8.03	7.66	7.51	8.88	9.18	8.63
CogVLM2	8.33	7.43	7.37	9.33	9.15	8.76
MiniGPT-4	6.71	6.43	5.98	6.72	7.34	5.93
Qwen3-VL	7.62	6.98	7.11	9.29	9.71	9.61

Table 8: Arrow format: per-judge scores and failure modes across all scaffolding conditions.

	Claude 4.5 Sonnet				**Gemini 2.5 Flash**				**Failure Modes (%)**				
	L2T	L2S	L3T	L3S	L2T	L2S	L3T	L3S	FC	NR	SO	CFS	UG
C1 (Full scaffolding)													
LLaVA-NeXT	7.31	7.40	6.61	6.72	8.46	7.55	7.47	6.73	99.50	0.30	0.00	0.05	0.15
LLaVA-RLHF	0.72	0.77	0.43	0.46	0.81	0.73	0.50	0.45	8.36	91.29	0.00	0.35	0.00
LLaVA-CoT	7.53	7.51	6.86	6.87	8.69	7.79	7.65	6.67	99.52	0.35	0.00	0.00	0.13
CogVLM2	5.60	4.44	4.69	3.45	6.37	3.17	4.71	2.25	48.53	9.98	16.34	24.92	0.23
MiniGPT-4	5.66	5.63	5.15	5.03	5.76	4.38	4.88	3.57	20.46	1.83	0.00	77.71	0.00
Qwen3-VL	7.68	7.70	6.99	7.17	9.26	8.88	8.69	8.41	99.22	0.78	0.00	0.00	0.00
C2 (Partial scaffolding)													
LLaVA-NeXT	7.12	7.20	6.53	6.42	7.73	7.43	7.27	6.78	98.45	0.55	0.00	0.92	0.08
LLaVA-RLHF	6.32	6.64	5.91	6.06	6.34	6.08	5.72	5.33	64.95	1.35	0.00	32.31	1.40
LLaVA-CoT	7.51	7.48	6.88	6.84	8.23	7.97	7.87	7.47	99.28	0.37	0.00	0.02	0.33
CogVLM2	5.89	3.81	5.12	3.06	6.89	4.55	5.79	3.43	13.23	7.23	21.41	56.84	1.30
MiniGPT-4	5.72	5.56	5.23	5.01	5.82	5.23	5.23	4.73	13.11	2.03	0.00	84.86	0.00
Qwen3-VL	7.76	7.78	7.14	7.35	8.43	8.38	8.23	8.13	98.63	1.37	0.00	0.00	0.00
C3 (Minimal scaffolding)													
LLaVA-NeXT	7.09	7.07	6.62	6.44	7.52	7.17	6.85	6.30	10.26	1.27	0.00	88.39	0.08
LLaVA-RLHF	6.40	6.66	6.17	6.38	6.36	5.57	5.93	5.51	4.56	2.23	0.00	93.05	0.15
LLaVA-CoT	7.50	7.43	6.95	6.90	8.06	7.76	7.59	7.16	79.93	0.92	0.08	15.84	3.23
CogVLM2	6.17	5.66	5.50	4.54	6.80	5.47	5.88	4.38	33.63	0.63	0.48	46.94	18.31
MiniGPT-4	6.08	6.15	5.71	5.66	6.35	5.89	5.75	5.30	0.58	3.75	0.00	95.67	0.00
Qwen3-VL	7.89	7.80	7.31	7.42	8.58	8.50	8.38	8.27	97.18	2.82	0.00	0.00	0.00

A.7 Human Evaluation

Expert human annotators evaluate responses from all six VLMs on 10 randomly selected images across all four formats and three scaffolding conditions. This yields 6,480 scored responses (10 images × 9 questions × 4 formats × 3 scaffolding × 6 VLMs, with text and structure assessed separately at L2 and L3).

Table 9: Semicolon format: per-judge scores and failure modes across all scaffolding conditions.

	Claude 4.5 Sonnet				Gemini 2.5 Flash				Failure Modes (%)				
	L2T	L2S	L3T	L3S	L2T	L2S	L3T	L3S	FC	NR	SO	CFS	UG
C1 (Full scaffolding)													
LLaVA-NeXT	7.40	7.50	6.77	6.96	8.08	7.92	7.81	7.60	99.73	0.22	0.00	0.00	0.05
LLaVA-RLHF	5.24	5.64	4.21	4.57	5.65	5.65	4.59	4.53	72.33	27.59	0.00	0.03	0.05
LLaVA-CoT	7.39	7.37	6.37	6.43	8.08	7.83	7.43	7.09	98.45	0.48	0.67	0.28	0.12
CogVLM2	6.08	5.23	5.22	4.16	6.92	5.27	5.78	4.15	66.37	3.56	10.93	18.62	0.52
MiniGPT-4	6.14	6.23	5.22	5.14	6.71	6.27	5.78	5.30	66.47	0.53	0.00	32.48	0.52
Qwen3-VL	7.86	7.87	7.18	7.39	8.56	8.51	8.35	8.27	99.35	0.65	0.00	0.00	0.00
C2 (Partial scaffolding)													
LLaVA-NeXT	7.14	7.11	6.56	6.36	7.80	7.39	7.44	6.80	99.20	0.42	0.00	0.08	0.30
LLaVA-RLHF	6.10	5.84	5.63	5.05	6.27	5.50	5.65	4.54	76.81	1.33	0.00	0.42	21.44
LLaVA-CoT	4.49	4.05	3.17	2.93	4.95	4.28	3.70	3.19	67.90	1.62	20.84	9.28	0.37
CogVLM2	5.01	2.11	4.54	1.44	6.07	2.84	5.28	2.01	20.84	9.20	23.55	40.95	5.46
MiniGPT-4	5.78	5.30	5.19	4.73	6.35	5.51	5.50	4.70	36.12	1.22	0.00	61.89	0.78
Qwen3-VL	7.85	7.87	7.20	7.42	8.48	8.44	8.29	8.18	98.87	1.13	0.00	0.00	0.00
C3 (Minimal scaffolding)													
LLaVA-NeXT	6.38	3.79	6.21	3.28	7.07	5.16	6.79	4.66	12.51	0.93	0.00	14.01	72.55
LLaVA-RLHF	4.99	2.55	4.89	2.95	4.75	2.37	4.35	2.09	14.21	1.47	0.00	29.89	54.44
LLaVA-CoT	6.49	6.57	5.17	5.00	7.09	6.54	5.83	5.16	85.09	10.89	0.40	0.78	2.89
CogVLM2	5.74	2.06	5.21	1.72	6.72	2.26	5.80	1.46	13.04	0.68	7.80	75.45	3.03
MiniGPT-4	6.15	5.20	5.60	4.68	6.73	5.56	6.10	4.88	0.83	1.63	0.00	96.55	0.98
Qwen3-VL	7.91	7.84	7.30	7.36	8.57	8.47	8.34	8.18	97.77	2.20	0.00	0.00	0.03

Table 10: Numbered format: per-judge scores and failure modes across all scaffolding conditions.

	Claude 4.5 Sonnet				Gemini 2.5 Flash				Failure Modes (%)				
	L2T	L2S	L3T	L3S	L2T	L2S	L3T	L3S	FC	NR	SO	CFS	UG
C1 (Full scaffolding)													
LLaVA-NeXT	7.59	7.72	6.84	7.14	8.12	8.11	7.82	7.72	99.65	0.35	0.00	0.00	0.00
LLaVA-RLHF	6.34	6.78	5.77	6.30	6.53	6.65	6.12	6.19	85.17	12.33	2.02	0.00	0.48
LLaVA-CoT	7.79	7.85	7.05	7.25	8.40	8.31	8.03	7.82	99.43	0.53	0.00	0.00	0.03
CogVLM2	6.81	4.26	6.23	4.06	5.82	3.99	5.39	3.88	55.39	6.96	34.62	0.00	3.03
MiniGPT-4	6.57	6.85	5.76	5.86	7.03	6.92	6.61	6.46	99.32	0.25	0.00	0.37	0.00
Qwen3-VL	7.99	8.03	7.40	7.70	8.55	8.55	8.42	8.38	99.13	0.87	0.00	0.00	0.00
C2 (Partial scaffolding)													
LLaVA-NeXT	7.48	7.62	6.96	7.16	7.93	7.84	7.58	7.29	98.90	1.07	0.00	0.03	0.00
LLaVA-RLHF	6.78	7.10	6.40	6.71	6.90	6.76	6.48	6.28	96.85	1.57	0.43	0.00	1.15
LLaVA-CoT	7.80	7.81	7.11	7.20	8.33	8.21	7.97	7.70	99.29	0.55	0.00	0.00	0.17
CogVLM2	5.64	3.53	5.01	2.81	6.42	3.76	5.48	2.81	45.59	8.15	43.86	0.00	2.40
MiniGPT-4	6.59	6.70	5.97	5.96	6.88	6.68	6.48	6.17	92.75	1.17	0.00	6.08	0.00
Qwen3-VL	8.08	8.06	7.53	7.76	8.58	8.62	8.47	8.46	98.50	1.50	0.00	0.00	0.00
C3 (Minimal scaffolding)													
LLaVA-NeXT	7.43	7.51	6.83	6.86	7.78	7.58	7.14	6.75	98.25	1.72	0.00	0.02	0.02
LLaVA-RLHF	6.51	6.62	6.14	6.14	6.38	5.98	5.80	5.23	87.81	1.87	0.03	0.02	10.28
LLaVA-CoT	7.73	7.66	7.13	7.20	8.07	7.94	7.68	7.43	97.25	0.88	0.00	0.00	1.87
CogVLM2	6.16	5.15	5.68	4.55	6.86	5.27	6.06	4.36	84.44	0.83	9.06	0.00	5.66
MiniGPT-4	6.66	6.88	5.98	5.78	6.91	6.53	6.32	5.69	97.68	2.08	0.00	0.13	0.10
Qwen3-VL	8.14	8.09	7.65	7.84	8.69	8.68	8.57	8.51	97.77	2.23	0.00	0.00	0.00

Tables 12–15 present human scores. Human evaluation confirms the key patterns from LLM-based evaluation: numbered format is scaffolding-robust (C1-to-C3 drops below 1 point for four models), the structural complexity boundary between linearized and graph-level formats holds (numbered-to-adjacency gap averages 3.5 points at C3), RLHF suppression at C1 is confirmed (arrow L3S = 0.40, adjacency L3S = 0.00), and Qwen3-VL maintains near-perfect robustness for linearized formats. The Pearson correlation between human and LLM judge ratings, computed on individual response scores, is 0.84 for structure and 0.58 for text. The strong structural correlation confirms that LLM judges reliably capture human assessments of causal structure quality. The moderate text correlation reflects greater subjectivity in natural language evaluation.

Table 11: Adjacency list format: per-judge scores and failure modes across all scaffolding conditions.

	Claude 4.5 Sonnet				Gemini 2.5 Flash				Failure Modes (%)				
	L2T	L2S	L3T	L3S	L2T	L2S	L3T	L3S	FC	NR	SO	CFS	UG
C1 (Full scaffolding)													
LLaVA-NeXT	6.97	6.89	6.05	5.73	7.58	7.14	6.79	5.91	93.99	0.23	0.00	5.58	0.20
LLaVA-RLHF	0.00	0.00	0.00	0.00	0.02	0.00	0.02	0.00	0.03	99.97	0.00	0.00	0.00
LLaVA-CoT	7.24	7.23	6.45	6.43	7.95	7.68	7.25	6.67	96.12	0.23	0.00	3.35	0.30
CogVLM2	4.14	0.77	4.08	0.55	4.99	1.61	4.87	1.43	11.14	25.57	24.99	38.08	0.22
MiniGPT-4	4.78	3.18	4.69	3.41	5.26	3.39	4.92	3.45	56.52	0.62	0.17	41.31	1.38
Qwen3-VL	7.73	7.91	6.96	7.20	8.45	8.39	8.09	7.90	97.88	0.73	0.00	1.37	0.02
C2 (Partial scaffolding)													
LLaVA-NeXT	5.88	4.30	5.92	4.19	6.78	5.18	6.44	4.46	39.28	0.62	0.00	16.79	43.31
LLaVA-RLHF	5.22	2.54	5.56	3.50	4.82	1.11	4.87	1.22	2.63	1.25	0.00	1.25	94.87
LLaVA-CoT	6.82	6.20	6.25	5.53	7.68	6.85	7.16	5.93	87.91	0.20	0.03	9.31	2.55
CogVLM2	4.61	0.31	4.75	0.24	5.45	1.58	5.44	1.63	6.73	12.98	29.12	48.43	2.73
MiniGPT-4	4.58	2.92	4.45	3.05	4.92	3.07	4.38	2.91	62.19	1.05	0.08	30.88	5.80
Qwen3-VL	6.95	6.67	6.33	5.90	8.08	7.69	7.54	6.92	82.91	1.62	0.00	15.29	0.18
C3 (Minimal scaffolding)													
LLaVA-NeXT	5.47	1.93	5.90	2.04	4.89	1.21	4.95	1.21	2.67	1.02	0.00	25.85	70.46
LLaVA-RLHF	4.79	0.41	4.77	0.33	4.41	0.99	4.13	0.80	0.00	1.72	0.00	49.44	48.85
LLaVA-CoT	5.60	4.17	5.34	4.22	6.45	5.04	6.01	4.71	71.60	1.00	0.92	22.76	3.73
CogVLM2	5.29	0.37	5.36	0.28	6.13	1.83	5.75	1.55	13.79	3.62	12.49	68.96	1.15
MiniGPT-4	4.48	1.89	4.50	2.06	4.49	1.84	4.26	1.94	48.39	1.63	5.31	30.00	14.66
Qwen3-VL	5.62	4.21	5.52	4.13	7.18	5.91	6.62	5.31	44.24	2.53	0.00	53.02	0.20

Table 12: Arrow format: human scores on 10 images across all scaffolding conditions.

Model	L1T	L2T	L2S	L3T	L3S
C1 (Full scaffolding)					
LLaVA-NeXT	7.97	7.43	7.03	7.50	6.30
LLaVA-RLHF	4.47	0.80	0.60	0.43	0.40
LLaVA-CoT	7.97	6.83	6.47	6.47	6.17
CogVLM2	8.63	4.97	2.23	5.20	3.07
MiniGPT-4	7.07	5.37	4.37	4.63	4.23
Qwen3-VL	8.87	8.03	8.00	7.43	7.27
C2 (Partial scaffolding)					
LLaVA-NeXT	7.33	7.66	7.16	7.53	7.00
LLaVA-RLHF	4.80	5.80	4.97	5.80	4.27
LLaVA-CoT	8.13	6.73	6.27	6.40	6.00
CogVLM2	8.87	4.90	2.70	5.00	2.47
MiniGPT-4	7.20	4.63	3.57	4.97	4.13
Qwen3-VL	9.07	7.90	7.80	7.57	7.37
C3 (Minimal scaffolding)					
LLaVA-NeXT	7.47	7.90	6.73	6.80	6.20
LLaVA-RLHF	4.37	6.20	5.50	5.60	4.80
LLaVA-CoT	8.20	6.50	5.87	6.23	5.67
CogVLM2	8.83	4.60	2.87	4.94	2.80
MiniGPT-4	7.27	5.33	4.07	5.60	4.03
Qwen3-VL	9.23	7.67	7.63	7.43	7.37

A.8 Complete CDD Values

Table 16 reports Causal Depth Disparity for all format-scaffolding conditions. CDD is computed on L3 scores averaged across judges: CDD $= (S_{\text{ST}}^{\text{text}} - S_{\text{ST}}^{\text{struct}})/S_{\text{ST}}^{\text{text}}$. Values marked with † indicate conditions where format compliance is below 50%; at these conditions, the structural score reflects the quality of whatever format the model actually produced (typically Cross-Format Substitution), and CDD should not be interpreted as a measure of the plausibility-faithfulness gap for the requested format.

Table 13: Semicolon format: human scores on 10 images across all scaffolding conditions.

Model	L1T	L2T	L2S	L3T	L3S
C1 (Full scaffolding)					
LLaVA-NeXT	7.53	7.66	7.36	7.00	6.80
LLaVA-RLHF	4.47	5.27	4.97	3.30	3.07
LLaVA-CoT	8.30	6.70	6.47	6.10	5.73
CogVLM2	8.47	4.50	4.13	5.00	3.60
MiniGPT-4	7.17	5.43	5.03	5.30	4.90
Qwen3-VL	9.07	7.60	7.50	7.27	7.10
C2 (Partial scaffolding)					
LLaVA-NeXT	7.70	7.60	7.73	6.40	6.96
LLaVA-RLHF	4.73	6.03	4.87	5.57	4.43
LLaVA-CoT	8.30	4.47	3.50	3.23	2.43
CogVLM2	8.50	4.57	1.10	4.13	1.43
MiniGPT-4	7.33	5.37	4.23	5.37	4.23
Qwen3-VL	9.13	7.57	7.47	7.33	7.10
C3 (Minimal scaffolding)					
LLaVA-NeXT	7.73	6.77	3.10	6.10	2.73
LLaVA-RLHF	5.07	4.33	3.10	3.60	2.33
LLaVA-CoT	8.23	6.57	6.31	5.67	5.07
CogVLM2	8.37	5.13	1.27	4.80	1.47
MiniGPT-4	7.17	5.27	3.80	5.33	3.80
Qwen3-VL	9.23	7.40	6.90	7.27	6.83

Table 14: Numbered format: human scores on 10 images across all scaffolding conditions.

Model	L1T	L2T	L2S	L3T	L3S
C1 (Full scaffolding)					
LLaVA-NeXT	7.77	7.73	7.23	7.37	7.03
LLaVA-RLHF	4.70	5.70	5.37	5.30	4.77
LLaVA-CoT	8.27	6.97	6.53	6.67	6.40
CogVLM2	8.77	4.43	2.90	4.97	2.73
MiniGPT-4	7.07	5.87	5.87	5.50	5.40
Qwen3-VL	9.17	7.47	7.37	7.00	6.93
C2 (Partial scaffolding)					
LLaVA-NeXT	7.77	7.83	7.33	7.07	6.53
LLaVA-RLHF	4.27	5.97	5.23	6.10	5.13
LLaVA-CoT	8.27	6.97	6.77	6.50	6.23
CogVLM2	8.63	3.80	1.77	3.50	1.87
MiniGPT-4	7.17	6.00	5.53	5.67	5.47
Qwen3-VL	9.20	7.30	6.97	7.03	6.93
C3 (Minimal scaffolding)					
LLaVA-NeXT	7.77	7.40	7.30	6.90	6.23
LLaVA-RLHF	4.43	5.83	5.33	5.57	5.10
LLaVA-CoT	8.27	6.80	6.53	6.40	6.33
CogVLM2	8.87	4.97	3.63	5.17	3.40
MiniGPT-4	7.27	5.93	5.90	5.53	5.13
Qwen3-VL	9.17	7.33	6.97	6.97	6.90

At high-compliance conditions (no †), CDD is near-zero for five of six models across numbered, arrow, and semicolon formats. This indicates that when models can produce the requested format, their structural output quality closely tracks their textual quality. CogVLM2 is the exception, retaining substantial CDD (0.24–0.46) even at accessible formats, consistent with its broad structural weakness.

Large CDD values (>0.3) co-occur almost exclusively with low compliance (†). Under these conditions, format production failure is the primary diagnostic signal, and CDD provides no additional information beyond what compliance already conveys.

Table 15: Adjacency List format: human scores on 10 images across all scaffolding conditions.

Model	L1T	L2T	L2S	L3T	L3S
C1 (Full scaffolding)					
LLaVA-NeXT	7.77	6.63	6.93	5.90	6.03
LLaVA-RLHF	4.17	0.00	0.00	0.00	0.00
LLaVA-CoT	8.27	6.70	6.53	6.03	5.77
CogVLM2	8.37	3.50	0.53	2.97	0.47
MiniGPT-4	7.30	3.97	3.47	4.07	3.50
Qwen3-VL	9.17	7.63	7.50	7.07	6.90
C2 (Partial scaffolding)					
LLaVA-NeXT	7.77	5.03	4.63	5.17	5.00
LLaVA-RLHF	4.57	5.00	2.00	4.77	2.00
LLaVA-CoT	8.23	6.07	5.47	5.93	5.33
CogVLM2	8.87	4.50	0.13	4.87	0.33
MiniGPT-4	7.33	4.03	3.30	3.50	3.20
Qwen3-VL	9.30	7.03	6.33	7.00	5.77
C3 (Minimal scaffolding)					
LLaVA-NeXT	7.80	5.43	2.07	5.60	2.03
LLaVA-RLHF	4.17	4.83	0.57	4.70	0.43
LLaVA-CoT	8.27	4.63	2.93	4.50	3.47
CogVLM2	8.43	5.07	0.47	5.43	0.40
MiniGPT-4	7.23	3.73	2.17	3.50	2.20
Qwen3-VL	9.17	5.97	3.80	5.37	3.60

Table 16: Complete CDD values (L3, averaged across judges). Near-zero = text and structure matched. † = compliance <50% (CDD not interpretable as a format-specific faithfulness measure).

		NeXT	RLHF	CoT	CogV2	Mini	Qwen
Num.	C1	−0.014	−0.050	+0.001	+0.317	+0.004	−0.016
	C2	+0.006	−0.009	+0.012	+0.464†	+0.026	−0.014
	C3	+0.026	+0.048	+0.012	+0.241	+0.067	−0.008
Arrow	C1	+0.045	+0.022†	+0.067	+0.394†	+0.143†	+0.006
	C2	+0.043	+0.021	+0.030	+0.405†	+0.069†	−0.007
	C3	+0.054†	+0.017†	+0.033	+0.216†	+0.044†	+0.000
Semi.	C1	+0.001	−0.034	+0.020	+0.245	+0.051	−0.008
	C2	+0.060	+0.150	+0.109	+0.649†	+0.118†	−0.007
	C3	+0.389†	+0.455†	+0.076	+0.711†	+0.183†	+0.006
Adj. L.	C1	+0.093	+0.000†	+0.044	+0.779†	+0.286	−0.003
	C2	+0.300†	+0.547†	+0.145	+0.816†	+0.325	+0.076
	C3	+0.700†	+0.873†	+0.213	+0.835†	+0.543†	+0.222†

A.9 Complete SSI Values

Table 17 reports Scaffolding Sensitivity Index and its decomposition for all formats at L3. $\text{SSI}_{\text{full}} = S_{\text{C1}}^{\text{struct}} - S_{\text{C3}}^{\text{struct}}$ measures total scaffolding sensitivity. $\text{SSI}_{\text{ex}} = S_{\text{C1}}^{\text{struct}} - S_{\text{C2}}^{\text{struct}}$ measures the effect of removing worked examples. $\text{SSI}_{\text{instr}} = S_{\text{C2}}^{\text{struct}} - S_{\text{C3}}^{\text{struct}}$ measures the effect of removing instructions. Positive values indicate degradation when scaffolding is removed.

Three patterns are salient. First, numbered format shows minimal scaffolding sensitivity across all models (SSI_{full} range: −0.49 to +0.57), confirming this format is scaffolding-robust. Second, LLaVA-RLHF shows large negative SSI_{ex} for arrows (−5.24) and adjacency lists (−2.36), meaning that removing worked examples substantially *improves* performance. This reflects suppression under full scaffolding (§4.3), not scaffolding independence. Third, adjacency list SSI_{full} is the largest positive value for four of six models (NeXT: +4.19, CoT: +2.10, Mini: +1.43, Qwen: +2.83), confirming that graph-level decomposition is the most scaffolding-dependent capability.

Table 17: Complete SSI decomposition (L3 structural scores, averaged across judges). Positive = degradation when scaffolding is removed; negative = improvement.

		NeXT	RLHF	CoT	CogV2	Mini	Qwen
Arrow	SSI_{full}	+0.35	−5.49	−0.26	−1.61	−1.18	−0.05
	SSI_{ex}	+0.12	−5.24	−0.38	−0.40	−0.57	+0.05
	SSI_{instr}	+0.23	−0.25	+0.12	−1.21	−0.61	−0.10
Semi.	SSI_{full}	+3.31	+2.03	+1.68	+2.57	+0.44	+0.06
	SSI_{ex}	+0.70	−0.24	+3.70	+2.43	+0.50	+0.03
	SSI_{instr}	+2.61	+2.27	−2.02	+0.14	−0.06	+0.03
Num.	SSI_{full}	+0.62	+0.56	+0.22	−0.49	+0.42	−0.13
	SSI_{ex}	+0.21	−0.25	+0.08	+1.16	+0.10	−0.07
	SSI_{instr}	+0.42	+0.81	+0.14	−1.65	+0.33	−0.07
Adj. L.	SSI_{full}	+4.20	−0.57	+2.08	+0.07	+1.43	+2.83
	SSI_{ex}	+1.50	−2.36	+0.82	+0.06	+0.45	+1.14
	SSI_{instr}	+2.70	+1.79	+1.27	+0.02	+0.98	+1.69

Table 18: LLaVA-CoT: baseline (B), single-format (SF), and multi-format multi-scaffolding (MF) fine-tuning. Per-judge L2 and L3 scores (Text / Structure) and failure modes (%).

	Claude 4.5 Sonnet				**Gemini 2.5 Flash**				**Failure Modes (%)**				
	L2T	L2S	L3T	L3S	L2T	L2S	L3T	L3S	FC	NR	SO	CFS	UG
Scaffolding robustness (semicolon)													
B C2	4.49	4.05	3.17	2.93	4.95	4.28	3.70	3.19	67.90	1.62	20.84	9.28	0.37
SF C2	7.87	7.83	7.37	7.61	8.61	8.45	8.42	8.28	99.87	0.13	0.00	0.00	0.00
MF C2	7.86	7.83	7.37	7.62	8.60	8.46	8.38	8.27	99.90	0.10	0.00	0.00	0.00
B C1	7.39	7.37	6.37	6.43	8.08	7.83	7.43	7.09	98.45	0.48	0.67	0.28	0.12
SF C1	7.82	7.82	7.22	7.51	8.53	8.38	8.29	8.15	99.95	0.05	0.00	0.00	0.00
MF C1	7.85	7.82	7.33	7.61	8.61	8.45	8.41	8.28	99.80	0.12	0.00	0.08	0.00
B C3	6.49	6.57	5.17	5.00	7.09	6.54	5.83	5.16	85.09	10.89	0.40	0.78	2.89
SF C3	7.85	7.81	7.36	7.61	8.60	8.43	8.39	8.25	99.83	0.10	0.07	0.00	0.00
MF C3	7.84	7.84	7.35	7.64	8.60	8.44	8.42	8.30	99.90	0.10	0.00	0.00	0.00
Format robustness (partial scaffolding)													
B Arrow	7.51	7.48	6.88	6.84	8.23	7.97	7.87	7.47	99.28	0.37	0.00	0.02	0.33
SF Arrow	7.65	7.53	6.95	7.03	8.47	8.22	8.06	7.79	1.87	0.13	0.00	98.00	0.00
MF Arrow	7.81	7.72	7.13	7.35	8.56	8.48	8.19	8.07	99.90	0.10	0.00	0.00	0.00
B Num.	7.80	7.81	7.11	7.20	8.33	8.21	7.97	7.70	99.29	0.55	0.00	0.00	0.17
SF Num.	7.58	7.14	7.11	7.16	8.41	7.93	8.13	7.78	0.15	0.22	0.00	99.64	0.00
MF Num.	7.92	7.94	7.57	7.88	8.65	8.64	8.50	8.49	99.92	0.08	0.00	0.00	0.00
B Adj.	6.82	6.20	6.25	5.53	7.68	6.85	7.16	5.93	87.91	0.20	0.03	9.31	2.55
SF Adj.	6.34	3.24	5.81	3.96	8.23	7.34	7.78	6.99	32.50	0.10	0.00	67.35	0.05
MF Adj.	7.82	7.92	7.52	7.97	8.61	8.63	8.26	8.08	99.67	0.05	0.00	0.20	0.08

A.10 Fine-Tuning Detailed Results

Both models are fine-tuned under two stages: single-format single-scaffolding (SF), using semicolon structures at partial scaffolding only, and multi-format multi-scaffolding (MF), using all four formats across all three scaffolding levels. All other training parameters are identical (§3.4; Appendix A.5). Tables 18 and 19 report per-judge scores and failure mode distributions for baseline (B), SF, and MF across all evaluation conditions.

LLaVA-CoT: Stage 1 analysis. The Cross-Format Substitution pattern under SF is uniform: 98.00% of arrow outputs and 99.64% of numbered outputs are produced in semicolon notation. The chain-of-thought architecture locks onto the trained format's surface notation regardless of the requested format. Despite producing the wrong format, causal content quality is maintained or slightly improved (arrow L3: 7.15 → 7.41; numbered: 7.45 → 7.47 averaged across judges). This confirms

Table 19: MiniGPT-4: baseline (B), single-format (SF), and multi-format multi-scaffolding (MF) fine-tuning. Per-judge L2 and L3 scores (Text / Structure) and failure modes (%).

	Claude 4.5 Sonnet				Gemini 2.5 Flash				Failure Modes (%)				
	L2T	L2S	L3T	L3S	L2T	L2S	L3T	L3S	FC	NR	SO	CFS	UG
Scaffolding robustness (semicolon)													
B C2	5.78	5.30	5.19	4.73	6.35	5.51	5.50	4.70	36.12	1.22	0.00	61.89	0.78
SF C2	7.11	7.19	6.39	6.56	8.02	7.71	7.52	7.20	96.64	0.17	1.18	2.02	0.00
MF C2	7.06	6.97	6.20	6.32	7.81	7.20	7.14	6.70	86.99	0.80	3.95	8.26	0.00
B C1	6.14	6.23	5.22	5.14	6.71	6.27	5.78	5.30	66.47	0.53	0.00	32.48	0.52
SF C1	6.47	4.38	6.10	5.00	7.43	4.68	7.11	5.33	61.10	0.23	38.67	0.00	0.00
MF C1	7.14	7.26	6.56	6.78	7.86	7.44	7.41	7.05	95.34	1.03	2.77	0.87	0.00
B C3	6.15	5.20	5.60	4.68	6.73	5.56	6.10	4.88	0.83	1.63	0.00	96.55	0.98
SF C3	5.99	1.29	5.67	1.62	7.25	2.84	6.82	3.09	14.88	0.43	84.69	0.00	0.00
MF C3	7.07	6.91	6.09	6.15	7.86	7.30	7.10	6.65	86.89	0.23	4.18	8.70	0.00
Format robustness (partial scaffolding)													
B Arrow	5.72	5.56	5.23	5.01	5.82	5.23	5.23	4.73	13.11	2.03	0.00	84.86	0.00
SF Arrow	6.91	6.51	6.20	5.99	7.79	7.12	7.18	6.71	86.06	0.23	13.71	0.00	0.00
MF Arrow	6.56	6.07	5.68	5.51	7.32	6.55	6.51	5.97	73.33	7.51	10.78	8.38	0.00
B Num.	6.59	6.70	5.97	5.96	6.88	6.68	6.48	6.17	92.75	1.17	0.00	6.08	0.00
SF Num.	7.10	7.15	6.45	6.55	7.84	7.61	7.30	7.09	95.22	0.15	1.17	3.46	0.00
MF Num.	7.18	7.49	6.80	7.17	7.87	7.93	7.46	7.50	99.05	0.32	0.48	0.15	0.00
B Adj.	4.58	2.92	4.45	3.05	4.92	3.07	4.38	2.91	62.19	1.05	0.08	30.88	5.80
SF Adj.	5.96	1.91	5.74	1.60	6.94	2.95	6.45	2.06	44.29	0.43	55.24	0.00	0.03
MF Adj.	5.86	2.57	5.35	2.14	7.33	4.43	6.33	3.01	54.34	2.00	43.65	0.00	0.00

that SF improved the model's causal reasoning while overwriting its format selection. Scaffolding robustness is near-perfect: compliance exceeds 99.83% across all three semicolon conditions.

LLaVA-CoT: Stage 2 analysis. MF resolves Cross-Format Substitution completely: CFS drops from 98.00% to 0.00% for arrows, from 99.64% to 0.00% for numbered, and from 67.35% to 0.20% for adjacency lists. Format compliance exceeds 99.67% across all four formats and all three semicolon scaffolding levels. Structural quality under MF exceeds SF for all non-semicolon formats: adjacency list L3 rises from 5.47 (SF) to 8.03 (MF averaged), numbered from 7.47 to 8.19, and arrow from 7.41 to 7.71. Both judges show consistent patterns: Gemini 2.5 Flash scores consistently higher (typical difference: 0.5–1.0 points), but relative rankings across conditions are preserved.

MiniGPT-4: Stage 1 analysis. The scaffolding collapse under SF follows a consistent pattern: Structure Omission is the dominant failure mode at non-trained scaffolding levels (84.69% at C3, 38.67% at C1). The model retains structural capability only at the scaffolding level it was trained on. Format transfer, by contrast, is strong for linearized formats: arrow compliance rises from 13.11% to 86.06% with 13.71% residual Structure Omission, and numbered compliance rises from 92.75% to 95.22%.

MiniGPT-4's L1 text quality improves from approximately 6.8 to 7.4 across all conditions after SF, including formats and scaffolding levels not used in training. At baseline, MiniGPT-4 shows negative AG for numbered format (requesting structure improves text quality). The fine-tuning result extends this pattern: structural training provides a general reasoning benefit for the standard projection architecture.

MiniGPT-4: Stage 2 analysis. MF substantially resolves the scaffolding collapse. Structure Omission at C3 drops from 84.69% (SF) to 4.18% (MF), and compliance rises from 14.88% to 86.89%. At C1, Structure Omission drops from 38.67% to 2.77%, with compliance rising from 61.10% to 95.34%. L3 quality at C3 improves from 2.35 (SF) to 6.40 (MF), a gain of +4.04. The scaffolding fragility that appeared architectural under SF was largely a training signal limitation.

Two residual format capacity limitations persist under MF. First, arrow compliance (73.33%) is lower than under SF (86.06%), with new failure modes appearing: 7.51% Complete Non-Response

and 10.78% Structure Omission. Under the multi-format multi-scaffolding protocol, each format-scaffolding combination receives approximately 2,500 training examples compared to 45,000 semicolons in Stage 1. The arrow compliance drop may partly reflect reduced per-format training dosage, and distinguishing this from a capacity ceiling would require further ablation. Second, adjacency list compliance improves only modestly (44.29% $\rightarrow$ 54.34%), with 43.65% Structure Omission persisting. This limitation is less attributable to training dosage: numbered format, which received the same proportional allocation, reached 99.05% compliance. Graph-level decomposition poses a qualitatively different challenge for the standard projection architecture.

The L1 text quality improvement observed under SF ($\sim$6.8 $\rightarrow$ $\sim$7.4) is preserved under MF ($\sim$7.4), confirming that the general reasoning benefit of structural training for MiniGPT-4 is independent of format diversity.

A.11 Limitations

Evaluation dataset scope. Our evaluation uses 1,000 MS-COCO images depicting natural scenes. Performance patterns may differ for synthetic images, domain-specific imagery (medical, satellite), or scenes with more complex causal structures than those typically present in everyday photographs.

Question generation. Causal questions are generated by GPT-4o, which may introduce biases in question difficulty or phrasing. Our manual quality check on 200 questions confirms >95% accuracy, but systematic biases affecting all questions uniformly cannot be ruled out. Since the same questions are used across all formats and scaffolding conditions, any such bias cannot explain the cross-format differences central to our findings.

Visual grounding of questions. Some causal questions, particularly at L2 and L3, may be answerable from world knowledge without visual grounding. Our Ungrounded Structure failure mode captures cases where models generate format-correct structures with generic placeholders in place of image-specific content. However, we do not include a systematic visual necessity verification for all questions. The cross-format variation in Ungrounded Structure rates (e.g., 72.55% for one model on semicolons vs. 0.08% on arrows at C3) indicates even the grounding dimension is format-contingent, but disentangling visual grounding failures from format production failures remains an open challenge.

Prompt asymmetry at minimal scaffolding. The adjacency list format requires more specification than the three linearized formats, so even the minimal scaffolding prompt carries more instructional content. The full scaffolding condition (C1) provides the cleanest cross-format comparison. The finding that five of six models show their lowest structural scores for the adjacency list even at C1 indicates the difficulty is intrinsic to the format, not an artifact of prompt length.

Scaffolding decomposition. The transition from C2 to C3 removes both the Pearl framework reference and the detailed procedural instructions simultaneously. We cannot isolate the individual contribution of each component.

LLM-as-judge limitations. Automated scoring by Claude 4.5 Sonnet and Gemini 2.5 Flash may not align with human judgment on all dimensions. We mitigate this with dual judges from different model families and human validation on 6,480 responses. Calibration differences exist between judges, but relative model rankings are preserved.

Fine-tuning scope. Targeted fine-tuning is applied to two of six models on one shared blind spot. The architectural asymmetry finding is established by this pair, but whether the pattern extends to other architectures and blind spots remains to be tested. MF inference was conducted for semicolon at all three scaffolding levels and for the remaining formats at partial scaffolding only. Scaffolding robustness after MF training for non-semicolon formats was not evaluated. For MiniGPT-4 under multi-format multi-scaffolding training, the drop in arrow compliance (86.06% to 73.33%) could reflect either a capacity ceiling or reduced per-format training dosage. Distinguishing these would require further ablation.

Model selection and capability range. Our evaluation covers six VLMs spanning five training approaches across a 2023–2025 capability range, from projection-based architectures (MiniGPT-4) to instruction-tuned models (Qwen3-VL). This range was chosen to test the framework across varying levels of instruction-following capability. Frontier models with stronger instruction following may show reduced format sensitivity for linearized formats. The specific compliance gaps we report are bound to the tested capability range and should not be extrapolated to stronger models without further testing. The methodology and score-compliance pairing, however, are applicable regardless of model scale. Qwen3-VL, the most capable model in our set, already shows near-zero format sensitivity for linearized formats but retains a 53-point compliance gap on adjacency lists, suggesting the structural complexity boundary may persist even as instruction following improves.

Causal specificity. Our evaluation tests format and scaffolding robustness for causal structure generation but does not include a non-causal control condition (e.g., generating sequential structures for non-causal content such as timelines or recipes). Format-specific fragility could partly reflect a general structured generation limitation, not one specific to causal reasoning. Disentangling causal-specific fragility from general format production difficulty would require matched causal and non-causal generation tasks. This would also test whether the evaluation design principles proposed in our discussion generalize beyond causal reasoning.