R-CARE: RELIABILITY-AWARE CONCEPT GRAPH CONDITIONING AND EVIDENCE VERIFICATION FOR VISUAL FIGURATIVE LANGUAGE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual figurative language understanding requires multimodal models to reason beyond the literal content of images and text by capturing implicit relations among visual cues, linguistic concepts, and figurative meanings. However, existing methods remain limited in structuring cross-modal figurative relations and in mitigating noise and semantic bias in automatically generated auxiliary knowledge. Moreover, explanations that are fluent and highly probable under the model may still be inconsistent with the multimodal evidence in the input. To address these challenges, we propose R-CARE (Reliability-aware Concept graph Adaptive Rhetorical Reasoning with Evidence Verification), a framework for visual figurative language reasoning. R-CARE first constructs a structured rhetorical concept graph that explicitly organizes visual entities, textual concepts, and potential figurative meanings. It then performs concept graph conditioned decoding, adaptively integrating graph structure and fine-grained semantic information based on the multimodal context while constraining their influence to mitigate interference from unreliable knowledge. Furthermore, R-CARE decouples label prediction from explanation verification. Given a fixed predicted label, it generates multiple candidate explanations, evaluates their consistency with the multimodal evidence using the model's internal multimodal representations, and selects the most evidence-consistent candidate subject to a generation-likelihood constraint. On the V-FLUTE benchmark, R-CARE achieves F1@0, F1@53, and F1@60 scores of 89.74, 78.14, and 60.85, respectively, surpassing MAPPER by 5.20, 8.45, and 9.13 percentage points. Code is available at: https://github.com/dahfs/R-CARE

1 INTRODUCTION

Multimodal large language models have made substantial progress in jointly understanding visual and textual information (Liu et al., 2023; Bai et al., 2023). Yet figurative phenomena such as metaphor, sarcasm, idiom, and humor pose a distinct challenge: understanding what an image "depicts" does not necessarily reveal what it "expresses" (Saakyan et al., 2025). Interpreting such non-literal meanings often requires reasoning over implicit figurative relations between visual content and textual concepts, rather than merely recognizing explicit multimodal cues (Chakrabarty et al., 2023; Zhang et al., 2023).

Existing studies improve visual figurative language understanding through zero-/few-shot reasoning with large vision-language models, task-specific fine-tuning, and auxiliary information such as scene graphs and image descriptions (Chen et al., 2024; Yang & Wang, 2024; Xu et al., 2024). These efforts show that enriched visual semantics can facilitate modeling of non-literal image–text relations (Kulkarni et al., 2024). However, figurative understanding often depends less on isolated entities or descriptions than on structured associations among visual concepts, textual concepts, and their figurative meanings (Zhang et al., 2024). Existing auxiliary representations mainly capture visible content and local relations, leaving such cross-modal figurative associations implicit and requiring models to integrate dispersed semantic cues during generation. Moreover, automatically generated auxiliary knowledge is vulnerable to errors, semantic bias, and hallucinations, limiting its reliability for downstream reasoning (Zechen et al., 2024; Huang et al., 2024).

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.

Although generative multimodal models can produce fluent and coherent explanations, generation quality does not guarantee consistency with the multimodal evidence underlying their predictions (Atanasova et al., 2023; Parcalabescu & Frank, 2024). Since explanations are typically generated autoregressively based on token-level likelihoods, high-probability outputs mainly reflect the model's learned generative preferences rather than their grounding in the current image–text evidence (Siegel et al., 2024; Reich et al., 2023). Consequently, even plausible explanations consistent with the predicted label may omit critical evidence or introduce unsupported inferences (Whitehouse et al., 2023; Madsen et al., 2024).

To address these challenges, we propose R-CARE (Reliability-aware Concept graph Adaptive Rhetorical Reasoning with Evidence Verification), a framework for reliable visual figurative language reasoning through structured knowledge organization, adaptive knowledge integration, and evidence-based explanation verification. R-CARE first constructs a rhetorical concept graph from the image description and claim, explicitly linking visual entities, textual concepts, and potential figurative meanings. It then employs reliability-aware concept graph conditioned decoding to adaptively integrate graph structure and semantics while limiting the influence of noisy or context-inconsistent knowledge. Finally, R-CARE decouples label prediction from explanation verification: given a fixed prediction, it generates multiple explanations and selects the candidate most consistent with the multimodal evidence under a generation-likelihood constraint. Our main contributions are summarized as follows:

- We propose R-CARE, a reliability-aware framework for visual figurative language understanding. R-CARE explicitly organizes cross-modal figurative relations through a structured rhetorical concept graph and adaptively controls the contribution of structured knowledge to reasoning based on the current multimodal context.
- We introduce a decision-preserving endogenous multimodal evidence verification mechanism that decouples label prediction from explanation verification. Given a fixed predicted label, the mechanism evaluates the consistency between candidate explanations and multimodal evidence, improving the evidence grounding of the generated explanations without altering the task decision.
- We conduct extensive experiments on the V-FLUTE benchmark. R-CARE achieves F1@0, F1@53, and F1@60 scores of 89.74, 78.14, and 60.85, respectively, outperforming MAPPER by 5.20, 8.45, and 9.13 percentage points.

2 Related Work

2.1 Knowledge-Augmented Visual Figurative Language Reasoning

Visual figurative language understanding requires integrating image–text cues to capture non-literal semantics, such as metaphor and sarcasm, together with their underlying conceptual associations (Zhang et al., 2021; Liang et al., 2022; Zhang et al., 2023). Recent studies improve this capability through auxiliary image descriptions, explanation supervision, and structured semantic representations (Liang et al., 2022; Chen et al., 2024; Yang & Wang, 2024; Mitra et al., 2024). For example, MAPPER supplements visual inputs with fine-grained image descriptions for relation prediction and explanation generation (Chen et al., 2024), while FigCLIP promotes figurative cross-modal alignment through representation fusion and explanation supervision (Yang & Wang, 2024). Beyond figurative language understanding, scene graphs have also been used to organize objects, attributes, and relations for compositional multimodal reasoning (Mitra et al., 2024). However, existing methods largely focus on semantic augmentation, cross-modal alignment, or generic structural information, without explicitly modeling figurative associations among visual entities, textual concepts, and potential meanings or accounting for the sample-specific reliability of automatically generated auxiliary knowledge. This motivates structured and reliability-aware knowledge modeling for visual figurative language reasoning.

2.2 Evidence-Grounded Multimodal Explanation Verification

Fluent and plausible explanations do not necessarily faithfully reflect the evidence underlying a model's decisions, making explanation faithfulness and evidence consistency central concerns in

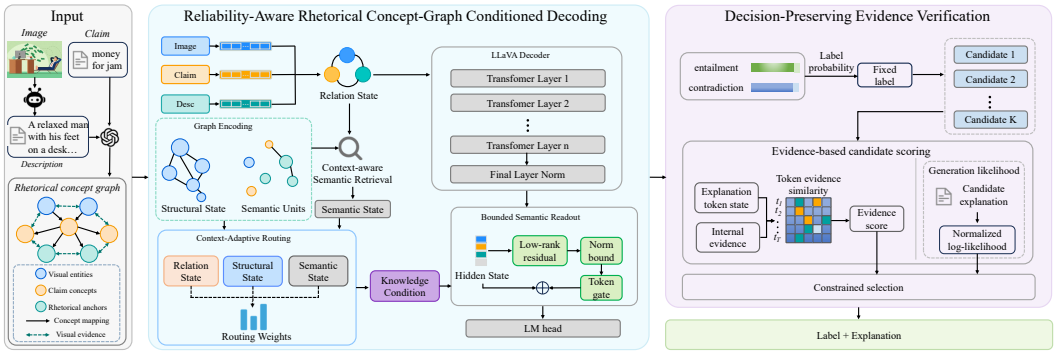


Figure 1: Overview of the proposed R-CARE framework.

explainable AI (Jacovi & Goldberg, 2020; Turpin et al., 2023; Madsen et al., 2024; Parcalabescu & Frank, 2024; Lyu et al., 2024). Existing studies assess this consistency through input interventions, counterfactual tests, attribution analysis, and self-consistency (Atanasova et al., 2023; Siegel et al., 2024; Kamahi & Yaghoobzadeh, 2024). Counterfactual editing and reconstruction test whether explanations capture prediction-relevant information (Atanasova et al., 2023), while attribution-, sufficiency-, and comprehensiveness-based metrics evaluate the faithfulness of vision-language explanations (Ambsdorf, 2023). CC-SHAP further compares input contributions to answer prediction and explanation generation, highlighting the distinction between explanation faithfulness and output-level self-consistency (Parcalabescu & Frank, 2024). Overall, these efforts shift explanation evaluation beyond surface-level quality toward consistency among explanations, model behavior, and input evidence. Building on this perspective, we use evidence consistency not only for evaluation but also for post-generation candidate selection, comparing explanations for a fixed prediction against model-internal multimodal evidence.

3 Method

3.1 Task Definition

Given an image–text pair (I, c), where I denotes the input image and c the claim to be verified, the model determines their semantic relation and generates a corresponding natural language explanation. Formally, the task is defined as

$$f_\theta(I, c) \rightarrow (y, e), \qquad y \in \mathcal{Y} = \{\text{entailment}, \text{contradiction}\}, \tag{1}$$

where y denotes the predicted relation label and e the generated explanation.

3.2 Framework Overview

As shown in Figure 1, R-CARE (Reliability-aware Concept graph Adaptive Rhetorical Reasoning with Evidence Verification) is built on LLaVA and improves generation reliability at two stages: knowledge incorporation and explanation verification. Given an image–claim pair, a visual interpreter first generates a claim-related image description, and ChatGPT then constructs a structured rhetorical concept graph from the description and claim to provide concept-level knowledge for subsequent reasoning. On the knowledge side, R-CARE employs reliability-aware concept graph conditioned decoding to adaptively integrate graph structure and fine-grained semantics according to the current multimodal context while constraining the influence of noisy or context-inconsistent knowledge. On the generation side, it decouples task prediction from explanation verification: after predicting the label, R-CARE keeps the prediction fixed, generates multiple candidate explanations, and selects the candidate most consistent with model-internal multimodal evidence.

3.3 Rhetorical Concept Graph Construction

Visual figurative language understanding requires modeling not only explicit visual and textual content, but also implicit associations between visual concepts and figurative meanings. To make these

relations explicit, R-CARE constructs a structured rhetorical concept graph $G = (V, E)$ from an image description and a claim, without using task labels or reference explanations. The node set V contains visual entities extracted from the image description, core textual concepts from the claim, and rhetorical anchors representing figurative semantics. The edge set E captures two types of relations: conceptual mappings between figurative expressions and their potential meanings, and visual-evidence links between textual concepts and corresponding visual entities. This graph organizes visual, textual, and figurative information into a unified cross-modal representation. We further derive graph-structure features and fine-grained semantic units from G for subsequent rhetorical-concept graph conditioned decoding. The complete graph-construction prompt, output schema, parsing rules, and JSON examples are provided in the appendix.

3.4 Reliability-Aware Rhetorical Concept Graph Conditioned Decoding (RACD)

Automatically constructed rhetorical concept graphs provide semantic relations beyond those explicitly expressed in the image and text, but may also contain erroneous, redundant, or context-irrelevant mappings. Incorporating such knowledge uniformly can therefore introduce noise into multimodal reasoning. To address this issue, R-CARE introduces reliability-aware rhetorical concept graph conditioned decoding (RACD), which adaptively selects and integrates graph knowledge according to the current multimodal context while constraining its influence on generation.

Given the representations of the image, claim, and image description, E_I, E_c, and E_d, we first project them into a shared semantic space and construct a cross-modal relation representation:

$$\begin{aligned} p_I = \phi_I(E_I), \qquad p_c = \phi_c(E_c), \qquad p_d = \phi_d(E_d), \\ r = [p_I; p_d; p_c; |p_d - p_c|; p_d \odot p_c]. \end{aligned} \tag{2}$$

A relation encoder then produces $h_r = f_r(r)$. For the rhetorical concept graph G, we encode its structural features g_{str} and fine-grained semantic units $\{e_j\}_{j=1}^{M}$ as

$$h_g = f_g(g_{\text{str}}), \qquad q_j = f_s(e_j), \quad j = 1, \ldots, M. \tag{3}$$

To identify graph semantics relevant to the current sample, we construct a context-dependent query from the cross-modal relation and graph-structure states and use it to aggregate semantic units:

$$u = f_q([h_r; h_g]), \qquad \beta_j = \text{softmax}_j\left(\frac{q_j^\top u}{\sqrt{d}}\right), \qquad h_s = \sum_{j=1}^{M} \beta_j q_j, \tag{4}$$

where h_s is the sample-specific graph-semantic state and d denotes the representation dimensionality.

We further learn context-adaptive routing weights to balance cross-modal relations, graph structure, and graph semantics:

$$\boldsymbol{\alpha} = \text{softmax}\left(f_{\text{route}}([h_r; h_g; h_s; g_{\text{str}}])\right), \qquad h_{\text{route}} = \alpha_r h_r + \alpha_g h_g + \alpha_s h_s, \tag{5}$$

where $\boldsymbol{\alpha} = [\alpha_r, \alpha_g, \alpha_s]$. The resulting knowledge condition is defined as

$$z = f_{\text{cond}}\left([h_r; h_g; h_s; h_{\text{route}}; |h_g - h_s|; h_g \odot h_s]\right). \tag{6}$$

This formulation allows the contribution of graph knowledge to vary across samples while explicitly modeling agreement between structural and semantic graph representations.

Rather than appending the concept graph as additional text tokens, RACD injects the knowledge condition z through a conditioned low-rank semantic readout after the final decoder normalization layer. For the hidden state h_t at decoding step t, we compute

$$a_t = \text{SiLU}(W_{\text{down}} h_t), \qquad m(z) = 1 + c_m \tanh(W_m z), \qquad \delta_t^{\text{raw}} = W_{\text{up}}\left(a_t \odot m(z)\right). \tag{7}$$

To prevent external knowledge from excessively perturbing the pretrained representation, we constrain the residual magnitude:

$$\bar{\delta}_t = \delta_t^{\text{raw}} \min\left(1, \frac{\eta \|h_t\|_2}{\|\delta_t^{\text{raw}}\|_2 + \epsilon}\right), \tag{8}$$

where η controls the maximum relative residual magnitude and $\epsilon > 0$ ensures numerical stability. Finally, a token-dependent gate controls the strength of knowledge injection:

$$\gamma_t = \gamma_{\max}\sigma\left(b + w_z^\top z + w_h^\top h_t\right), \qquad \widetilde{h}_t = h_t + \gamma_t \bar{\delta}_t, \tag{9}$$

where c_m controls channel-wise modulation and $\gamma_{\max}$ bounds the token-level gating strength. RACD therefore combines sample-adaptive knowledge selection with constrained token-level modulation, allowing useful rhetorical knowledge to influence decoding while limiting interference from unreliable graph information.

3.5 Decision-Preserving Endogenous Multimodal Evidence Verification (DPEV)

Although reliability-aware decoding mitigates interference from noisy structured knowledge, high-likelihood explanations may still be weakly grounded in the multimodal evidence encoded by the model. Generation likelihood reflects a candidate's probability under the learned distribution, but does not directly measure its consistency with the current image, claim, and image description. To address this limitation, R-CARE introduces decision-preserving endogenous multimodal evidence verification (DPEV), which separates label prediction from explanation verification. DPEV first selects the label by comparing the conditional probabilities of the complete label sequences, then generates and verifies candidate explanations while keeping this label fixed. Decision preservation is defined relative to this explicitly selected label, which may differ from the label produced by ordinary joint generation.

Given the multimodal context $\mathcal{X}$, the model first predicts the task label:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p_\theta(y \mid \mathcal{X}). \tag{10}$$

Conditioned on the fixed prediction $\hat{y}$, we then generate K candidate explanations using beam search:

$$\mathcal{E}_{\hat{y}} = \{e^{(k)}\}_{k=0}^{K-1} = \mathrm{Beam}_K\left[p_\theta(e \mid \mathcal{X}, \hat{y})\right], \tag{11}$$

where $e^{(0)}$ denotes the default explanation. Subsequent verification only affects explanation selection and does not re-estimate the predicted label.

For each candidate $e^{(k)}$, we retain the decoding hidden states $h_{k,t}$ over its explanation-token region $\mathcal{T}_k$. The explanation states and internal representations of the image, claim, and image description are projected into a shared normalized evidence space:

$$v_{k,t} = \mathrm{Norm}(\psi_h(h_{k,t})), \qquad r_m = \mathrm{Norm}(\psi_{\mathrm{role}}(E_m)), \quad m \in \{I, c, d\}. \tag{12}$$

We first measure token-level similarity to each evidence source and aggregate it using a length-normalized smooth maximum:

$$a_{k,t,m} = v_{k,t}^\top r_m, \qquad c_{k,m} = \tau\left[\log \sum_{t \in \mathcal{T}_k} \exp\Big(\frac{a_{k,t,m}}{\tau}\Big) - \log|\mathcal{T}_k|\right], \tag{13}$$

where τ is the aggregation temperature. The resulting $c_{k,m}$ emphasizes strongly aligned explanation tokens while reducing sensitivity to candidate length.

We further capture global explanation–evidence consistency and token-level evidence association:

$$\bar{v}_k = \mathrm{Norm}\left(\frac{1}{|\mathcal{T}_k|}\sum_{t \in \mathcal{T}_k} v_{k,t}\right), \qquad p_{k,m} = \bar{v}_k^\top r_m, \qquad u_k = \frac{1}{|\mathcal{T}_k|}\sum_{t \in \mathcal{T}_k} \max_m a_{k,t,m}. \tag{14}$$

Let $\mathbf{c}_k = [c_{k,I}; c_{k,c}; c_{k,d}]$ and $\mathbf{p}_k = [p_{k,I}; p_{k,c}; p_{k,d}]$. We combine local coverage, global matching, and cross-source coverage statistics into an evidence feature vector:

$$\xi_k = \Big[\mathbf{c}_k; \mathbf{p}_k; \min_m c_{k,m}; \mathrm{mean}_m\, c_{k,m}; \mathrm{std}_m\, c_{k,m}; u_k\Big], \qquad S_k = f_{\mathrm{ev}}(\xi_k), \tag{15}$$

where S_k is a relative evidence-consistency score used only for candidate comparison rather than as a calibrated probability.

To avoid selecting evidence-consistent but unlikely explanations, we additionally compute the length-normalized generation likelihood:

$$\ell_k = \frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \log p_\theta \left(e_t^{(k)} \mid e_{<t}^{(k)}, \mathcal{X}, \hat{y} \right). \tag{16}$$

Using the default explanation $e^{(0)}$ as the reference, a candidate is eligible only if it is valid, preserves the predicted label, improves evidence consistency by a margin τ_e, and remains sufficiently likely under the generation model:

$$\mathcal{A} = \left\{ k \in \{1, \dots, K-1\} \;\middle|\; \mathrm{valid}(e^{(k)}),\ \mathrm{label}(e^{(k)}) = \hat{y},\ e^{(k)} \neq e^{(0)},\ S_k > S_0 + \tau_e,\ \ell_k \geq \ell_0 - \tau_\ell \right\}. \tag{17}$$

Finally, DPEV selects the eligible candidate with the highest evidence score; if no candidate satisfies the constraints, the default explanation is retained:

$$\hat{e} = \begin{cases} e^{(k^\star)}, & k^\star = \arg\max_{k \in \mathcal{A}} S_k,\ \mathcal{A} \neq \emptyset, \\ e^{(0)}, & \mathcal{A} = \emptyset. \end{cases} \tag{18}$$

3.6 Training Objectives and Staged Optimization

R-CARE is trained in stages. We first jointly optimize the LoRA parameters and the knowledge-conditioned controller to adapt the generator to rhetorical concept graph conditioning. We then freeze the LoRA parameters and further refine the controller. Finally, the generator and graph-conditioning modules are fixed, and only the conditioned semantic readout and semantic alignment projections are optimized.

During semantic readout refinement, the objective is

$$\mathcal{L}_{\mathrm{RCD}} = \mathcal{L}_{\mathrm{DRO}} + \lambda_y \mathcal{L}_{\mathrm{lbl}} + \frac{\lambda_d}{2} \left(w_g \mathcal{L}_g + w_l \mathcal{L}_l^{\mathrm{rel}} \right), \tag{19}$$

where $\mathcal{L}_{\mathrm{DRO}}$ denotes the GroupDRO-reweighted autoregressive generation loss, and $\mathcal{L}_{\mathrm{lbl}}$ is the label prediction loss. $\mathcal{L}_g$ encourages the global explanation representation to align with the target explanation, while $\mathcal{L}_l^{\mathrm{rel}}$ enforces reliability-weighted local concept coverage according to the consistency of each concept with the target explanation and current knowledge condition. Gradients are stopped through the reliability weights, and w_g and w_l are adaptively adjusted according to the historical scales of the two semantic losses. Target explanations are used only as supervision during training and are not provided at inference time.

After knowledge-conditioned decoding is trained, we freeze the generator and graph-conditioning modules and optimize only the endogenous multimodal evidence verifier. Given a positive explanation e^+ and a negative explanation e^- that shares the same task label but mismatches the multimodal evidence, with evidence scores S^+ and S^-, respectively, the verifier objective is

$$\mathcal{L}_{\mathrm{EV}} = T\ \mathrm{softplus}\left(\frac{m - S^+ + S^-}{T} \right) + \frac{\lambda_a}{2} \left[\mathrm{softplus}(-S^+) + \mathrm{softplus}(S^-) \right]. \tag{20}$$

where m is the ranking margin and T the temperature coefficient. The first term encourages S^+ to exceed S^- by a margin, while the second stabilizes their score scales. Because e^+ and e^- share the same task label, the verifier is encouraged to distinguish explanations by multimodal evidence consistency rather than label cues.

4 Experiments

4.1 Dataset

We use V-FLUTE as the main evaluation benchmark. Each sample contains an image, a figurative claim, a binary relation label, and a natural language explanation, covering diverse figurative phenomena such as metaphors, similes, idioms, sarcasm, and humor across varied visual styles. We follow the official data split, with 4,578 training, 726 validation, and 723 test samples.

4.2 BASELINES

We compare R-CARE with various baselines, including fine-tuned LLaVA variants Liu et al. (2023), zero-shot LLaVA-1.6 Liu et al. (2024), API-based vision-language models, InstructBLIP-7B Dai et al. (2023), and GPT-4o Hurst et al. (2024). In model names, VF denotes fine-tuning on V-FLUTE; Image denotes an ablation where the real image is replaced with a white square; eViL denotes fine-tuning on e-SNLI-VE or e-ViL Do et al. (2020); Kayser et al. (2021); ZS, SG, and 5-shot denote zero-shot, scene-graph prompting, and five-shot inference strategies, respectively. MAPPER results are recomputed using the same implementation Chen et al. (2024).

4.3 EVALUATION METRICS

We follow the automatic evaluation protocol of Saakyan et al. and report F1@0, F1@53, and F1@60. F1@0 is the standard label F1, evaluating only prediction correctness. F1@53 and F1@60 additionally penalize samples whose ExplanationScore falls below the respective threshold, even when the predicted label is correct. ExplanationScore is computed as the mean of BERTScore and BLEURT between the generated and reference explanations. Together, these metrics evaluate both label accuracy and explanation quality.

4.4 EXPERIMENTAL SETUP

We use LLaVA-v1.6-Vicuna-7B as the base multimodal model for R-CARE. Visual descriptions are generated by the frozen base model, while ChatGPT 5.6 Sol is used to construct rhetorical concept graphs as structured external knowledge. For parameter-efficient adaptation, we apply LoRA with rank $r = 128$, scaling factor $\alpha = 256$, dropout 0.05, and a maximum sequence length of 2048. LoRA training runs for 3 epochs with a learning rate of 2×10^{-4} and a warmup ratio of 0.03, while the internal evidence verifier is trained for 2 epochs. Beam search uses $K = 3$ candidates. The base model is quantized to 4-bit NF4 with BF16 computation, and all experiments are conducted on a single 48GB NVIDIA A40 GPU.

4.5 MAIN RESULTS

Table 1 presents the main results on the V-FLUTE test set. R-CARE achieves the best performance across all three metrics, reaching 89.74, 78.14, and 60.85 on F1@0, F1@53, and F1@60, respectively. Compared with MAPPER, R-CARE improves these metrics by 5.20, 8.45, and 9.13 percentage points, and outperforms GPT-4o 5-shot by 10.32, 8.79, and 4.54 points, respectively. These results demonstrate consistent improvements in both relation prediction and explanation-aware evaluation.

Notably, the gains over MAPPER are larger on F1@53 and F1@60 than on F1@0, with a 9.13-point improvement on the stricter F1@60 metric. Since F1@53 and F1@60 require both a correct relation prediction and an explanation meeting the corresponding quality threshold, these gains suggest that R-CARE improves not only label prediction but also explanation quality. This pattern is consistent with the design of reliability-aware concept graph conditioned decoding and endogenous evidence verification, which aim to reduce interference from unreliable knowledge and favor evidence-consistent explanations.

Table 1: Main results of different methods on V-FLUTE. Bold indicates the best result, and underline indicates the second-best result.

Model	F1@0	F1@53	F1@60
InstructBLIP-7B-SG	38.03	4.15	1.38
InstructBLIP-7B-ZS	43.37	4.14	2.07
LLaVA-1.6-ZS-7B	45.44	35.57	18.38
Random Baseline	49.82	–	–
LLaVA-1.6-ZS-7B + SG	52.94	39.27	14.86
Gemini-1.5-Pro	53.70	39.72	19.01
LLaVA-1.5-7B-eViL	54.34	4.11	0.55
LLaVA-1.6-ZS-34B	55.60	48.32	31.83
Claude-3 Opus	56.07	45.37	22.31
LLaVA-1.6-ZS-34B + SG	58.08	45.74	26.77
GPT-4	64.00	56.22	38.56
LLaVA-1.5-7B-Image	64.77	53.28	39.37
Gemini-1.5-Pro 5-shot	67.25	56.04	37.14
Claude-3 Opus 5-shot	67.79	58.70	35.32
GPT-4 5-shot	69.36	61.95	49.81
LLaVA-1.5-7B-VF	72.78	60.66	47.12
LLaVA-1.5-7B-eViL+VF	74.91	62.34	48.80
GPT-4o	75.41	60.97	37.20
GPT-4o 5-shot	79.42	69.35	<u>56.31</u>
MAPPER	<u>84.54</u>	<u>69.69</u>	51.72
R-CARE	**89.74**	**78.14**	**60.85**

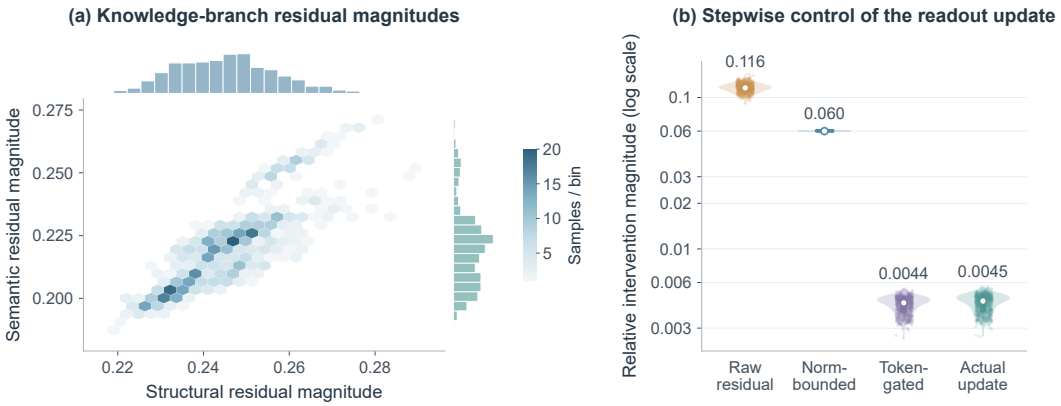


Figure 2: Internal behavior of RACD in knowledge conditioning and update control.

4.6 Ablation Study

As shown in Table 2, we analyze the contributions of visual descriptions, RACD, and DPEV to R-CARE. Removing either DPEV or RACD consistently degrades performance across all three metrics. Without DPEV, F1@0, F1@53, and F1@60 decrease from 89.74, 78.14, and 60.85 to 88.58, 76.46, and 59.74, respectively, while removing RACD causes larger drops to 86.78, 74.39, and 57.62. These results indicate that both modules contribute to performance, with RACD providing a larger gain by introducing and regulating structured rhetorical knowledge during reasoning.

Table 2: Ablation study of R-CARE components. CD denotes the claim-conditioned image description generated based on the input claim.

Model	F1@0	F1@53	F1@60
R-CARE (Full)	**89.74**	**78.14**	**60.85**
w/o DPEV	88.58	76.46	59.74
w/o RACD	86.78	74.39	57.62
w/o Both (LLaVA+CD)	85.10	71.36	54.05
LLaVA-SFT	80.91	67.90	49.27

When both modules are removed, performance further decreases to 85.10, 71.36, and 54.05. Relative to this variant, adding RACD improves F1@0, F1@53, and F1@60 by 3.48, 5.10, and 5.69 points, while adding DPEV improves them by 1.68, 3.03, and 3.57 points. The larger gains on F1@53 and F1@60 suggest that both components are particularly beneficial for explanation-aware evaluation. In addition, the improvement of LLaVA + CD over LLaVA-SFT shows that visual descriptions provide useful supplementary visual semantics for subsequent reasoning.

4.7 Analysis of the Internal Behavior of RACD

As shown in Figure 2(a), the structural and semantic branches exhibit clear sample-dependent variation across the 723 test samples, with mean residual magnitudes of 0.246 and 0.220 and a strong positive correlation (Pearson $r = 0.836$), indicating coordinated rather than fixed knowledge conditioning across samples. Figure 2(b) further shows that the median relative magnitude of the unconstrained low-rank residual decreases from 0.116 to 0.060 after norm clipping, which is activated for 98.33% of explanation tokens, while token-level gating further reduces the update magnitude to 0.0044; the final applied update remains at 0.0045, approximately 3.82% of the original residual magnitude. These results show that RACD combines sample-adaptive knowledge conditioning with constrained token-level updates, allowing structured knowledge to influence generation while limiting excessive perturbation of the pretrained decoding states.

4.8 Analysis of the Behavior and Effectiveness of DPEV

As shown in Figure 3(a), DPEV jointly considers evidence gain and generation likelihood when selecting among non-default explanations. A candidate is eligible only when $\Delta S > 0.05$ and $\Delta \ell \geq -0.10$, after which the candidate with the highest evidence score is selected. Among 1,441 scorable non-default candidates from 723 test samples, DPEV replaces the default explanation for

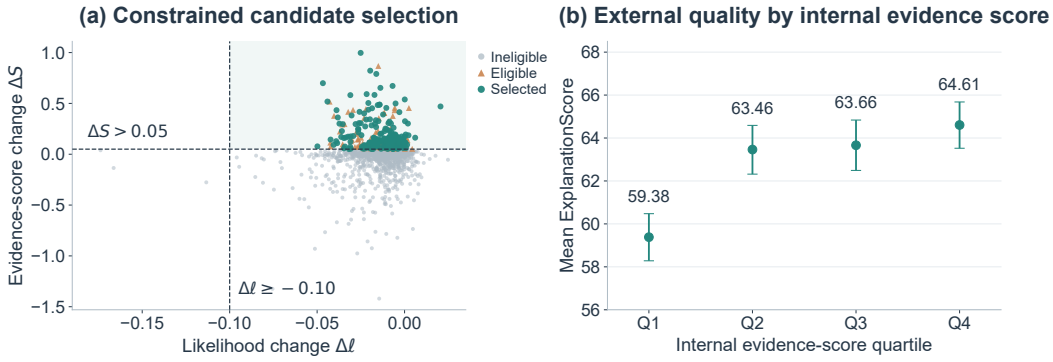


Figure 3: Selection behavior and external validity of DPEV.

204 samples while leaving all predicted labels unchanged, demonstrating selective explanation revision under decision preservation. Figure 3(b) further examines the relationship between internal evidence scores and externally evaluated explanation quality. Dividing the 723 default explanations into four score quantiles yields mean ExplanationScore values of 59.38, 63.46, 63.66, and 64.61 from Q1 to Q4, respectively. The internal evidence score is also positively correlated with ExplanationScore (Spearman $\rho = 0.214$, 95% bootstrap CI $[0.144, 0.283]$). Although this score is not a calibrated measure of explanation quality, these results suggest that it captures multimodal evidence signals associated with external explanation quality and can therefore support evidence-based explanation selection without altering the original task decision.

4.9 Case Analysis

Figure 5 (see Appendix C) presents a representative V-FLUTE idiom example comparing how different models connect visual evidence with figurative meaning. For "shrinking violet," MAPPER identifies the idiom as referring to shyness but interprets the child's face-covering behavior mainly as surprise or worry, leading to an incorrect relation prediction. In contrast, R-CARE associates this visual cue with shyness or timidity and links it to the non-literal meaning of the idiom, yielding a reasoning path consistent with the reference semantics. DPEV further compares candidate explanations while preserving the predicted label and selects the one more consistent with the multimodal evidence. As a result, the final explanation more explicitly captures the intended meaning of "shrinking violet" as "a very shy person." This case illustrates how R-CARE improves the alignment between local visual cues, figurative semantics, and the generated explanation.

5 Conclusion

This paper presents R-CARE, a reliability-aware framework for visual figurative language reasoning that addresses the structured modeling of cross-modal figurative relations, the reliability of automatically generated auxiliary knowledge, and the evidence consistency of generated explanations. R-CARE organizes visual entities, textual concepts, and potential figurative meanings into rhetorical concept graphs, and introduces reliability-aware concept graph conditioned decoding to adaptively incorporate structured knowledge while limiting its influence on generation. It further employs decision-preserving endogenous multimodal evidence verification to select evidence-consistent explanations without altering the predicted label. On V-FLUTE, R-CARE achieves F1@0, F1@53, and F1@60 scores of 89.74, 78.14, and 60.85, outperforming MAPPER by 5.20, 8.45, and 9.13 percentage points, respectively. These results demonstrate the effectiveness of reliability-aware structured knowledge integration and model-internal evidence verification for multimodal figurative reasoning and explanation generation.

Reproducibility Statement

We provide detailed model definitions, training objectives, and inference procedures for R-CARE. Experiments use the official V-FLUTE split and evaluation protocol, with all major training and inference hyperparameters, quantization settings, and hardware configurations reported in Section 4.4. The mathematical formulations and optimization procedures of RACD and DPEV are specified in Section 3. The appendix further provides the complete rhetorical concept graph construction prompt, structured output schema, parsing rules, and an illustrative JSON example. These details are intended to support reproducibility of both graph construction and the subsequent training and inference pipeline.

AI Use Statement

Generative AI tools were used in this work in two capacities. First, ChatGPT 5.6 Sol was used as part of the research pipeline to construct rhetorical concept graphs from image descriptions and textual claims, as described in the main text and appendix. Second, generative AI was used to assist with language editing and improving the clarity of the manuscript. All methodological decisions, implementations, experiments, results, and AI-assisted content were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

References

Jakob Ambsdorf. Benchmarking faithfulness: towards accurate natural language explanations in vision-language tasks. *arXiv preprint arXiv:2304.08174*, 2023.

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. Faithfulness tests for natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 283–294, 2023.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7370–7388, 2023.

Jiale Chen, Qihao Yang, Xuelian Dong, Xiaoling Mao, and Tianyong Hao. A textual modal supplement framework for understanding multi-modal figurative language. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pp. 85–91, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. e-snli-ve: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*, 2020.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. Visual hallucinations of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9614–9631, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4198–4205, 2020.

Sepehr Kamahi and Yadollah Yaghoobzadeh. Counterfactuals as a means for evaluating faithfulness of attribution methods in autoregressive language models. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 452–468, 2024.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1244–1254, 2021.

Shreyas Kulkarni, Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. A report on the figlang 2024 shared task on multimodal figurative language. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pp. 115–119, 2024.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1767–1777, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 50(2):657–723, 2024.

Andreas Madsen, Sarath Chandar, and Siva Reddy. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 295–337, 2024.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14420–14431. IEEE, 2024.

Letitia Parcalabescu and Anette Frank. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6048–6089, 2024.

Daniel Reich, Felix Putze, and Tanja Schultz. Measuring faithful and plausible visual grounding in vqa. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 3129–3144, 2023.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. Understanding figurative meaning through explainable visual entailment. In *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. 1–23, 2025.

Noah Siegel, Oana-Maria Camburu, Nicolas Heess, and Maria Perez-Ortiz. The probabilities also matter: A more faithful metric for faithfulness of free-text explanations in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 530–546, 2024.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Chenxi Whitehouse, Tillman Weyde, and Pranava Swaroop Madhyastha. Towards a unified model for generating answers and explanations in visual question answering. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1693–1705, 2023.

Yanzhi Xu, Yueying Hua, Shichen Li, and Zhongqing Wang. Exploring chain-of-thought for multimodal metaphor detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 91–101, 2024.

Qihao Yang and Xuelin Wang. Figclip: A generative multimodal model with bid yosef ectional cross-attention for understanding figurative language via visual entailment. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pp. 92–98, 2024.

Bai Zechen et al. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. Multimet: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 3214–3225, 2021.

Dongyu Zhang, Jingwei Yu, Senyuan Jin, Liang Yang, and Hongfei Lin. Multicmet: A novel chinese benchmark for understanding multimodal metaphor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6141–6154, 2023.

Linhao Zhang, Jintao Liu, Li Jin, Hao Wang, Kaiwen Wei, and Guangluan Xu. Gome: Grounding-based metaphor binding with conceptual elaboration for figurative language illustration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18500–18510, 2024.

A Rhetorical Concept Graph Construction: Prompt and Illustrative Example

A.1 Graph-construction Prompt

We construct a rhetorical concept graph from the image description and textual claim. The graph contains visual entities, claim concepts, and rhetorical anchors, together with concept mappings and links to visual evidence. Graph construction does not use task labels, reference explanations, or model predictions. The inferred_phenomenon field is inferred solely from the wording of the claim, while the dataset phenomenon field is left empty. The prompt used for graph construction is provided below.

```
Given an image description and a textual claim, construct a rhetorical
concept graph.  Extract visual entities, core claim concepts, and
figurative expressions, and explicitly connect rhetorical concepts
to visual evidence.  Return the specified fields as valid JSON.
Distinguish metaphor, simile, and idiom using the claim and image
description alone.  Do not use dataset labels, reference explanations,
or model predictions.
```

A.2 Illustrative example

The following example illustrates the rhetorical concept graph constructed from an image description and its corresponding claim. The graph explicitly organizes visual entities, claim concepts, figurative anchors, conceptual mappings, and visual-evidence links, while relying only on the image description and claim without using task labels or reference explanations.

```
{
  "id": 217,
  "image": "image/test/217.png",
  "claim": "The ground is as dry as a desert",
  "description": "The image shows a close-up of cracked earth with a
    rough texture. The cracks are irregular and deep, creating a
    pattern that fills the entire frame. The color palette is muted,
    with shades of brown and gray dominating the scene. The image is
    taken from a top-down perspective, focusing on the texture and
    pattern of the cracks.",
  "phenomenon": "",
  "inferred_phenomenon": "simile",
  "visual_nodes": [
    "cracked earth",
    "deep, irregular cracks",
    "rough, brown-gray surface"
  ],
  "claim_nodes": [
    "ground",
    "dryness",
    "desert",
    "extreme dryness"
  ],
  "figurative_anchors": [
    "as dry as a desert"
  ],
  "concept_mappings": [
    {
      "source": "as dry as a desert",
      "target": "extreme dryness",
      "relation": "simile",
      "meaning": "The ground is explicitly compared with a desert in
     terms of dryness."
    }
  ],
  "evidence_edges": [
    {
      "source": "ground",
      "target": "cracked earth",
      "type": "visual_support"
    },
    {
      "source": "extreme dryness",
      "target": "deep, irregular cracks",
      "type": "visual_support"
    },
    {
      "source": "as dry as a desert",
      "target": "deep, irregular cracks",
      "type": "anchor_to_visual"
    }
  ]
}
```


B Effect of Candidate Budget

Figure 4 analyzes the effect of the candidate budget K on prediction performance and inference cost. As shown in Figure 4(a), F1@0 remains unchanged at 89.74 across all settings, confirming that varying the number of candidate explanations does not affect the predicted task label. Increasing K from 1 to 3 steadily improves the explanation-aware metrics: F1@53 rises from 76.75 to 78.14, while F1@60 increases from 57.11 to 60.85. Further increasing K to 5 yields only a marginal

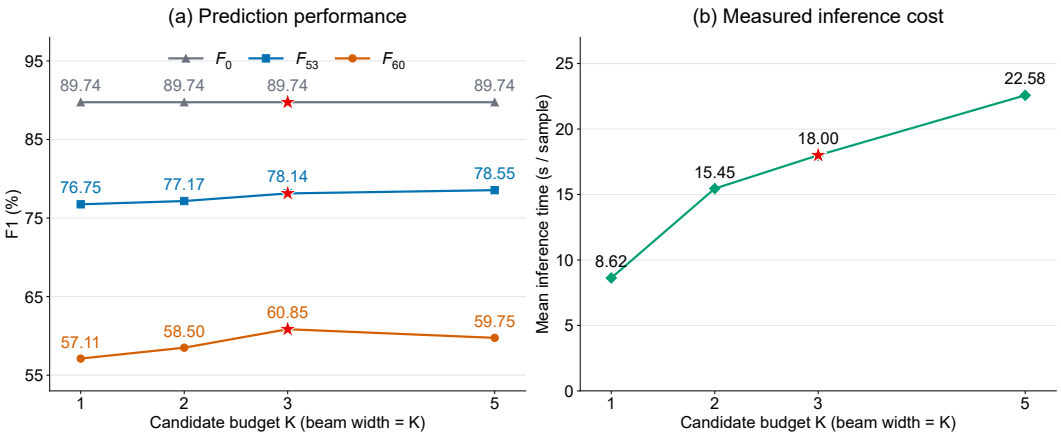


Figure 4: Effect of candidate budget K on prediction performance and inference cost.

Input	MAPPER	R-CARE
	Dataset: IRFL (idiom) Claim: "shrinking violet" Gold label: Entailment Visual description: The image features a young child with their hands covering their eyes. The child has light-colored hair and is wearing a striped shirt. The child's eyes are visible, and the focus is on the eyes and the hands covering them. The background is dark, which contrasts with the child's light-colored hair and shirt.	
Figurative Grounding	Face-covering gesture ↓ Startle or concern ↓ Mismatch with shyness ↓ Contradiction	Face-covering gesture ↓ Shyness or timidity Timidity associated with violets C0 S=0.862 / Timidity associated with a shrinking violet C1 S=0.877 / a very shy person C2 S=0.935 Shrinking violet ↓ Entailment
Predicted label	Contradiction	Entailment
ExplanationScore	60.75	70.23
Model output	contradiction. The image depicts a young child with a startled or concerned expression, with one hand partially covering the face. The metaphor "shrinking violet" suggests someone who is timid or introverted, which contradicts the image because the child's expression does not convey timidity but rather a more intense or emotional state.	entailment. The image depicts a young child with their hands over their eyes, appearing shy or timid. The metaphor "shrinking violet" is entailed by the image because the child's demeanor and the way they are hiding their eyes are reminiscent of the timidity often associated with a shrinking violet, a metaphor for a person who is very shy.
Reference explanation	The image depicts a child covering most of their face with their hands, seemingly bashful or shy. This visual suggests a reticence or timidness that relates to the idiom "shrinking violet," which refers to a person who is very shy or reluctant to draw attention. The child's demeanor embodies the essence of the idiom, thus the image entails the concept of being a shrinking violet.	

Figure 5: Case analysis on a V-FLUTE idiom example.

gain on F1@53 (78.55) and reduces F1@60 to 59.75. Figure 4(b) shows that the mean inference time increases from 8.62 s/sample at $K = 1$ to 18.00 s/sample at $K = 3$ and 22.58 s/sample at $K = 5$. These results indicate that a moderate candidate budget provides sufficient diversity for evidence-based explanation selection, whereas further enlarging the search space introduces additional computation with limited or inconsistent performance gains. We therefore use $K = 3$ in the main experiments as a balance between explanation quality and inference efficiency.

C Additional Case Analysis

Figure 5 provides the complete case visualization, including the input, figurative grounding process, model outputs, and evidence-based candidate selection.

Table 3: Performance of R-CARE under different training seeds on V-FLUTE.

Method / Run	Seed Pair	F1@0	F1@53	F1@60
MAPPER	–	84.54	69.69	51.72
R-CARE (Main)	(42, 2026)	89.74	78.14	60.85
R-CARE (Run 2)	(123, 123)	89.33	77.86	58.49
R-CARE (Run 3)	(456, 456)	89.68	77.87	60.03
R-CARE (Mean $\pm$ Std.)	–	89.58 ± 0.22	77.95 ± 0.16	59.79 ± 1.20
Gain over MAPPER	–	$+5.04$	$+8.26$	$+8.07$

D Sensitivity to Random Seeds

Table 3 examines the sensitivity of R-CARE to training randomness using the main run and two additional random seed configurations. All three runs consistently outperform MAPPER across F1@0, F1@53, and F1@60. Averaged across the three runs, R-CARE achieves 89.58, 77.95, and 59.79, exceeding MAPPER by 5.04, 8.26, and 8.07 percentage points, respectively. The relatively small standard deviations on F1@0 (0.22) and F1@53 (0.16) indicate stable relation prediction and explanation-aware performance across these runs, while F1@60 exhibits higher variation (1.20), suggesting greater sensitivity under the stricter explanation-quality threshold. These results provide an additional robustness check across training seeds, although the limited number of runs is not intended to support formal statistical significance testing.

E Human Evaluation

To complement the automatic metrics, we conduct a human evaluation of R-CARE-generated explanations. We sample 99 test instances using stratified sampling across the six phenomenon–dataset groups listed in Table 4, covering metaphors and similes, idioms, sarcasm, and humor. Two annotators independently evaluate all instances without access to each other's judgments.

Evaluation Protocol. Each explanation is rated on a five-point scale according to four criteria: (1) correctness of the predicted relation label; (2) consistency with the intended figurative meaning; (3) support from the image, image description, and claim; and (4) semantic consistency with the human-written reference explanation. Scores from 1 to 5 correspond to incorrect or unsupported, mostly incorrect, partially correct, mostly correct, and fully correct explanations, respectively. Scores of 4–5 are regarded as adequate, whereas scores of 1–3 are regarded as inadequate. When the predicted label contradicts the gold label, the score is generally capped at 2 unless the reference annotation is ambiguous.

Score Aggregation and Agreement. Let $s_{ij} \in \{1, \ldots, 5\}$ denote the score assigned to sample i by annotator j. We average the two ratings and linearly rescale the resulting score to the range $[0, 100]$:

$$H_i = \frac{\left(\frac{1}{2}\sum_{j=1}^{2} s_{ij}\right) - 1}{4} \times 100. \tag{21}$$

For each stratum g, we report $\bar{H}_g = \frac{1}{|g|}\sum_{i \in g} H_i$. The 95% confidence intervals are estimated using 20,000 sample-level bootstrap resamples. To assess inter-annotator reliability, we convert the five-point ratings into binary adequacy judgments and compute observed agreement and Cohen's κ. The annotators achieve 85.86% observed agreement, with Cohen's $\kappa = 0.718$ and a 95% bootstrap confidence interval of $[0.584, 0.839]$.

Results. As shown in Table 4, R-CARE achieves an overall human evaluation score of 63.01, with a 95% bootstrap confidence interval of $[56.19, 69.57]$. Performance is highest on the IRFL metaphor/simile subset (91.35), whereas NYCartoons (42.86) and HAIVMET (49.11) are more challenging, suggesting greater difficulty with context-dependent visual humor and visually grounded metaphorical associations. The two annotators assign mean raw scores of 3.38 and 3.66,

Table 4: Human evaluation of R-CARE generated explanations across figurative phenomena and source datasets.

Phenomenon	Dataset	N	Human Avg. (0–100)	95% Bootstrap CI
Metaphor/Simile	HAIVMET	14	49.11	[30.36, 68.75]
Metaphor/Simile	IRFL (metaphor/simile)	13	91.35	[75.00, 100.00]
Idioms	IRFL (idiom)	17	62.50	[45.59, 77.94]
Sarcasm	MuSE	14	69.64	[55.36, 81.25]
Humor	MemeCap	27	63.89	[50.93, 75.93]
Humor	NYCartoons	14	42.86	[32.14, 55.36]
Overall	All	99	63.01	[56.19, 69.57]

respectively, and 93.94% of paired ratings differ by no more than one point. These results indicate generally consistent judgments, while the difference in mean scores suggests some variation in annotator strictness. Because several strata contain relatively few samples, subset-level differences should be interpreted together with their confidence intervals.