

RIFT: Relation Modeling and Transfer for Visual Figurative Entailment

Xiaojian Lou, Qimeng Yang*, Hui Song, Yunjing Jiang, Jinmiao Song, Xinjun Pei
Xinjiang University, Urumqi, China
yangqm@xju.edu.cn

Abstract

Visual figurative entailment requires determining whether an image supports the nonliteral meaning of a textual claim, posing challenges for multimodal figurative understanding. Existing methods typically use fixed image-text fusion and struggle to capture sample-specific relations between visual evidence and figurative claims. Moreover, the limited scale of annotated visual figurative data makes it difficult to learn robust figurative relations from visual supervision alone. To address these challenges, we propose RIFT, which generates a claim-conditioned visual description and uses a relation-gated multimodal mixture of experts to model interactions among the image, description, and claim. It further retrieves figurative reasoning patterns from a FLUTE-based relation-level rationale memory and employs a confidence routing gate to modulate this external guidance according to retrieval reliability. The resulting internal and external relation representations are integrated as latent prefix tokens for joint entailment prediction and explanation generation. On V-FLUTE, RIFT achieves 88.52 F1@0, 76.07 F1@53, and 59.47 F1@60, outperforming MAPPER by 3.98, 6.38, and 7.75 points, respectively. These improvements, particularly on explanation-constrained metrics, show that relation modeling and cross-modal transfer benefit both prediction and explanation quality.

1 Introduction

Figurative language is fundamental for expressing abstract concepts, affective attitudes, and intentions during communication and is widely used in social media, news, advertising, comics, and everyday communication (Lakoff and Johnson, 2008; Shutova, 2011; Troiano et al., 2018). With the

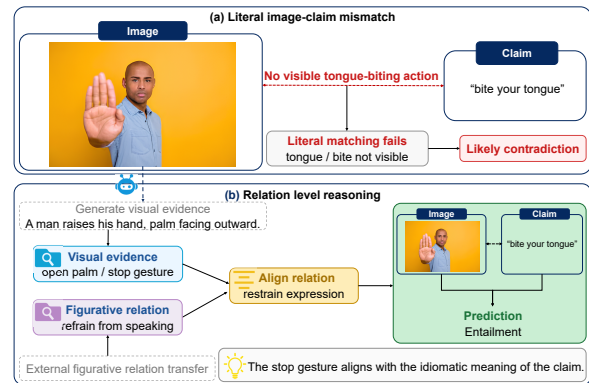


Figure 1: An illustrative example of visual figurative entailment with the idiom “bite your tongue.” (a) A model relying on literal matching finds no explicit visual evidence for tongues or bites and incorrectly predicts *contradiction*. (b) Claim-conditioned visual evidence and external figurative relation transfer help identify that the stop gesture aligns with the idiomatic meaning of restraining expression and correctly predict *entailment*.

proliferation of multimodal content on the web, figurative understanding has increasingly extended beyond text to image-text scenarios. Visual figurative entailment presents a particularly challenging problem, in which a model must go beyond recognizing literal objects in an image and determine whether the depicted actions, scenes, or visual symbols support the nonliteral meaning of a textual claim.

Existing methods often fuse image and text features into a unified representation, implicitly assuming that a single fusion pattern suffices for modeling diverse image-claim interactions (Li et al., 2023; Dai et al., 2023; Ye et al., 2024). However, visual figurative entailment requires models to adaptively capture how visual and textual cues jointly convey nonliteral meaning within each image-claim pair. For instance, idiomatic claims may rely on symbolic actions, sarcastic claims may hinge on semantic tension between image and text, and metaphors or similes may involve abstract attribute mapping

*Corresponding author. Code is available at: <https://github.com/constant-lou/RIFT>

(Kulkarni et al., 2024). Without such adaptive modeling, models risk overemphasizing isolated objects or surface-level lexical cues, making it difficult to determine whether an image supports the intended nonliteral meaning. As shown in Figure 1(a), the image offers relational cues such as an open palm gesture, yet a model relying primarily on literal matching may incorrectly predict “contradiction” due to the absence of explicit visual evidence for words like tongue or bite. This illustrates that fixed fusion strategies are insufficient for capturing latent nonliteral relations within each image-claim pair.

On the other hand, current visual figurative entailment methods mainly learn figurative reasoning from limited image-text annotations but rarely explicitly use transferable relation-level explanations from textual figurative entailment datasets (Saakyan et al., 2025; Chen et al., 2024). Such explanations provide more than semantically similar sentences; they encode reusable reasoning patterns, such as action symbol-communicative intention, visual cue-idiomatic meaning, and semantic conflict-sarcasm judgment (Chakrabarty et al., 2022). Without an understanding of these external relations, models can only implicitly determine these patterns on the basis of scarce image-text samples, which may lead to insufficient or incorrect reasoning in difficult cases where literal evidence is absent. As shown in Figure 1(b), correctly identifying the claim “bite your tongue” as entailment requires more than detecting a concrete object, namely, a hand. The model must understand that, in an idiomatic context, “bite your tongue” conveys the meaning of restraining one’s expression and that the visual action of an open palm or a stop gesture aligns with this meaning at an abstract semantic level. Therefore, visual figurative entailment requires access to reusable relation-level reasoning patterns that can complement the limited supervision available in image-text annotations.

To address these two challenges, we propose relation modeling and transfer for visual figurative entailment (RIFT). RIFT first employs a relation-gated multimodal mixture of experts (MoE) strategy to model sample-level relations among the image, the claim-conditioned description, and the claim. This module allows the model to adaptively emphasize different relationships among visual cues, the generated descriptions, and figurative claims for each sample. Then, a relation-level rationale memory is constructed from FLUTE, and textual explanations whose relation patterns are rel-

evant to the current sample are retrieved. To reduce the influence of unreliable retrievals, RIFT introduces a confidence routing gate that dynamically controls the injection of external rationales on the basis of the retrieval confidence. By combining the current visual evidence with transferable relation patterns, RIFT not only improves nonliteral semantic understanding but also enables the generation of more complete and reliable explanations.

The main contributions of this paper are as follows:

- We identify a dual-relation requirement for visual figurative entailment: the adaptive modeling of sample-specific cross-modal relations and the transfer of reusable reasoning patterns from external figurative resources.
- We propose RIFT, a relation-centric framework that integrates internal relation modeling with external figurative relation transfer. RIFT employs a relation-gated multimodal MoE strategy to model relations among the image, description, and claim, and retrieves reusable figurative reasoning patterns from a relation-level rationale memory constructed from FLUTE.
- We evaluate RIFT on V-FLUTE, where it outperforms MAPPER by 3.98, 6.38, and 7.75 points in F1@0, F1@53, and F1@60, respectively. Ablation studies further verify the effectiveness of each proposed component, with notable improvements on explanation-constrained metrics.

2 Related Work

2.1 Multimodal Figurative Reasoning and Cross-Modal Fusion

Recent studies have explored explicit cross-modal modeling for multimodal figurative reasoning. MAPPER adopts a describer-thinker framework that generates image-text supplements and combines descriptions with claims for prediction and explanation (Chen et al., 2024). FigCLIP builds on CLIP and GPT-2, using bidirectional fusion and cross-attention to improve image-text alignment for figurative understanding (Yang and Wang, 2024). ImaRA leverages conceptual metaphor theory to model cross-domain associations for low-resource metaphor detection and explanation (Tian et al.,

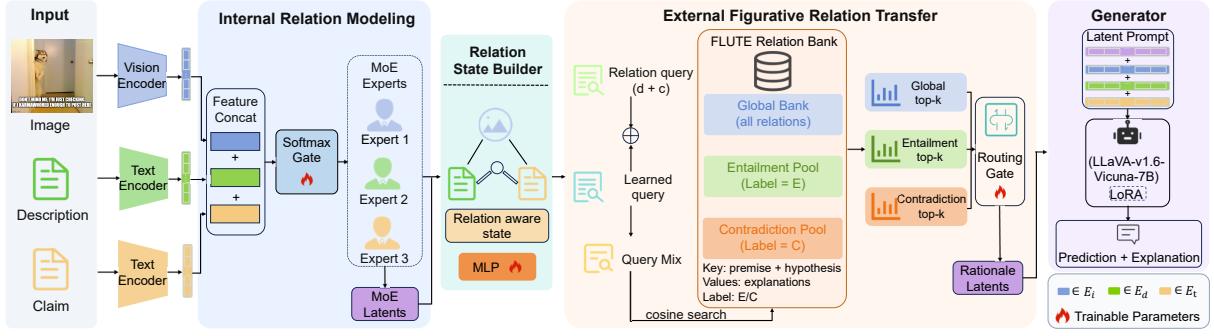


Figure 2: Overview of RIFT. The internal relation modeling branch encodes the image, claim-conditioned description, and claim, routing the concatenated features through a softmax gate across specialized MoE experts to produce relation-aware latent tokens. The external figurative relation transfer branch retrieves top- K rationales from the global, entailment, and contradiction views of the FLUTE relation bank via a mixed query. A confidence routing gate then controls the injection strength of external rationale latents. The resulting internal and external representations are injected into LLaVA as a unified latent prefix for joint prediction and explanation generation.

2025). These studies confirm that multimodal figurative reasoning requires abstract relation modeling beyond simple object recognition.

2.2 Reasoning-Oriented Adaptation for Multimodal Models

Recent work has enhanced multimodal reasoning through explicit reasoning supervision and efficient parameter adaptation. Multimodal-CoT separates rationale generation from answer inference to introduce chain-of-thought reasoning into multimodal tasks (Zhang et al., 2023), while ScienceQA provides multimodal QA data with lectures and explanations to support visual and knowledge-grounded reasoning (Lu et al., 2022). Reinforcement learning approaches such as Perception-R1 further improve reasoning via perception-oriented rewards that encourage stronger visual grounding (Xiao et al., 2025). Efficient adaptation methods including LoRA and prefix tuning demonstrate that low-rank updates or continuous prefix representations can effectively adapt pretrained models to downstream tasks (Hu et al., 2021; Li and Liang, 2021).

3 Method

The overall architecture of RIFT is shown in Figure 2. RIFT formulates visual figurative entailment as an internal-external relation reasoning process. The internal branch jointly encodes the image, claim-conditioned visual description, and textual claim, routing the concatenated features through a softmax gate to multiple experts, each capturing complementary relation patterns, to produce MoE latent tokens. The external branch retrieves rationales from FLUTE most relevant to the current sample,

providing transferable nonliteral reasoning information. A confidence routing gate then controls the injection strength based on retrieval confidence, label direction, and relation representation, reducing interference from low-relevance retrievals while preserving reliance on visual evidence. Finally, both representations are organized into a unified latent prefix and injected into LLaVA, fine-tuned via LoRA to jointly generate the entailment prediction and explanation.

3.1 Task Definition

Given an image-text sample (I, c) , where I denotes the input image and c denotes the textual claim to be verified, the goal is to determine the semantic relation between I and c and generate a corresponding natural language explanation. The model predicts a binary label as follows:

$$y \in \{\text{entailment}, \text{contradiction}\} \quad (1)$$

Simultaneously, an explanation e is generated in the following unified output format:

$$[\text{Prediction}]. [\text{Explanation}] \quad (2)$$

Unlike standard visual entailment, visual figurative entailment requires that the model not only recognize literal objects in the image but also determine whether the depicted actions, scenes, and visual symbols support the nonliteral meaning of the claim through relation-level cues.

3.2 Generation of the Claim-Conditioned Visual Descriptions

In multimodal figurative entailment, the semantic relation between an image and a claim often involves nonliteral mapping. Relying solely on raw

visual encoder features may be insufficient for capturing claim-relevant visual cues. Therefore, we use LLaVA-v1.6-Vicuna-7B as a visual interpreter and prompt it to generate a claim conditioned visual description for each sample (I, c) , with the prompt details provided in Appendix.

The generated description d helps expose visual cues in the image that are most relevant to the claim c , serving as textual signals for downstream multimodal relation modeling. The parameters of the visual interpreter are kept frozen during the generation of the description, and no reference labels or explanations are used. The main model subsequently takes the triplet (I, d, c) as input for the multimodal relation inference and explanation generation steps.

3.3 Internal Relation Modeling

Given an image I , the claim-conditioned description d , and the textual claim c , RIFT encodes the three input streams separately and maps them to a shared latent space, yielding image, description, and claim representations. These representations are concatenated into a multimodal relation context component x_{full} , which is fed into both a gating network and multiple relation experts.

Unlike fixed fusion strategies, RIFT employs a relation-gated multimodal MoE strategy to dynamically model the internal relations of the current sample. Each expert is implemented as a 2-layer MLP with independent parameters, receiving the same complete relation context, thereby encouraging different experts to learn complementary cross-modal relation transformations. The gating network assigns expert weights on the basis of the overall context of the current sample as follows:

$$w = \text{Softmax}(g(x_{\text{full}})),$$

$$Z_{\text{int}} = \sum_{m=1}^M w_m f_m(x_{\text{full}}) \quad (3)$$

where f_m denotes the m -th relation expert, w_m is its activation weight for the current sample, and M is the total number of experts. The resulting Z_{int} is converted to a set of internal relation tokens that capture relations among the image, description, and claim within the current sample.

3.4 External Figurative Relation Transfer

While the internal branch captures multimodal evidence within the current sample, visual figurative entailment data are limited in scale and may not

cover the diverse nonliteral reasoning patterns required for figurative understanding. In contrast, FLUTE provides textual figurative entailment instances with premises, hypotheses, labels, and explanations that encode transferable relation-level rationales. RIFT therefore builds a FLUTE relation memory and transfers these external reasoning patterns to the current visual sample via retrieval.

Relation Query Construction. For the current sample, RIFT constructs a textual relation query using the generated description d as the premise and the claim c as the hypothesis. This query targets rationales with similar figurative relation patterns rather than superficially similar sentences. To incorporate visual evidence and explicit description-claim interactions, the image, description, and claim representations are first projected into a shared latent space, yielding p_I , p_d , and p_c , respectively. RIFT then constructs an interaction-aware relation state:

$$x_{\text{rel}} = [p_I; p_d; p_c; |p_d - p_c|; p_d \odot p_c] \quad (4)$$

where $[\cdot; \cdot]$ denotes concatenation, $|p_d - p_c|$ captures the absolute discrepancy between the description and claim, and $\odot$ denotes element-wise multiplication. The relation state is mapped into the key space of the FLUTE relation memory via a trainable MLP, producing the normalized multimodal query:

$$q_{\theta} = \text{Normalize}(\text{MLP}(x_{\text{rel}})) \quad (5)$$

The textual query q_0 and multimodal query q_{θ} are then fused with a learnable weight:

$$q_{\text{mix}} = \text{Normalize}((1 - \alpha)q_{\theta} + \alpha q_0) \quad (6)$$

where α balances textual relation matching and multimodal evidence, enabling description-claim relations to guide retrieval while retaining visual information.

FLUTE Relation Memory. RIFT builds the external relation memory from the FLUTE training set. Each record contains a relation key, a rationale value, and a label. The relation key encodes the FLUTE premise-hypothesis pair for similarity retrieval; the rationale value encodes the corresponding natural language explanation to produce external relation tokens; the label indicates whether the rationale corresponds to entailment or contradiction.

To reduce interference between rationales of different label types, RIFT adopts three retrieval views: a global bank, an entailment pool, and a contradiction pool. The global bank captures overall relation similarity, while the entailment and contradiction pools provide label-specific reasoning patterns. Given q_{mix} , the model retrieves the top- K rationales from each view and aggregates their embeddings by retrieval similarity. The aggregated representations are passed through a rationale adapter-decoder comprising a projection layer and a token expansion layer, producing a fixed number of latent rationale tokens aligned with the LLaVA embedding dimension. External textual explanations are thus converted into continuous relation-level guidance tokens rather than being directly copied into the generator.

Confidence Routing Gate. Retrieved external rationales may not always match the current visual sample, and unconditional injection can introduce noise and weaken reliance on visual evidence. RIFT therefore introduces a confidence routing gate to estimate the reliability of external relation transfer for each sample.

A lightweight auxiliary classifier is applied to the internal relation representation Z_{int} to predict the entailment or contradiction probability p_y , supervised by the reference label via cross-entropy during training. During inference, p_y serves only as a routing signal. RIFT then computes the external injection weight by combining the current relation representation, retrieval similarities across views, the margin between entailment and contradiction retrieval branches, and the certainty of the auxiliary label:

$$\rho = \sigma(f_\rho([x_{\text{rel}}; s_g; s_{\text{route}}; \Delta; |2p_y - 1|])) \quad (7)$$

where x_{rel} is the current relation representation obtained by projecting Z_{int} , s_{route} is the label-based similarity of a particular view, Δ is the confidence margin between entailment and contradiction retrieval branches, and $|2p_y - 1|$ is the certainty of the auxiliary label.

A larger ρ strengthens the contribution of external relation tokens, while a smaller ρ suppresses them to reduce noise from low-relevance retrievals. The rationale adapter-decoder then produces external relation tokens scaled by ρ , yielding the final external relation representation Z_{ext} .

3.5 Latent Prefix Injection and Training

RIFT concatenates the internal relation tokens Z_{int} and external relation tokens Z_{ext} into a unified latent prefix prepended to the original LLaVA input embeddings:

$$X' = [Z_{\text{int}}; Z_{\text{ext}}; X_{\text{LLaVA}}] \quad (8)$$

where X_{LLaVA} denotes the original multimodal input embeddings. This prefix allows the generator to attend to both internal multimodal relations and external figurative reasoning patterns during generation.

To explicitly supervise the learned relation-aware representations, we introduce a query-alignment loss and an auxiliary label loss:

$$\begin{aligned} \mathcal{L}_q &= 1 - \cos(q_\theta, q_0), \\ \mathcal{L}_{\text{label}} &= \text{BCEWithLogits}(h(x_{\text{rel}}), y) \end{aligned} \quad (9)$$

where $h(\cdot)$ is the auxiliary label classifier, and $y \in \{0, 1\}$ is the reference entailment/contradiction label. The two auxiliary objectives are jointly optimized with the autoregressive language modeling objective.

The model autoregressively generates outputs in the format [prediction].[explanation], where [prediction] is either entailment or contradiction. During training, all components are jointly optimized end-to-end except the frozen FLUTE relation memory and visual interpreter, including the relation modeling modules, confidence routing gate, and LoRA parameters. The top- K retrieval operations are excluded from gradient backpropagation to maintain retrieval stability. The final training objective combines the autoregressive language modeling loss with the contrastive loss, the MoE balancing loss, the query-alignment loss $\mathcal{L}_q$, and the auxiliary label loss $\mathcal{L}_{\text{label}}$.

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. The dataset covers multiple figurative phenomena, including metaphors, similes, idioms, sarcasm, and humor, across diverse visual styles, enabling systematic evaluation of cross-modal figurative reasoning and explanation generation.

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	56.31
MAPPER	84.54	69.69	51.72
RIFT (Ours)	88.52	76.07	59.47

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

We follow the official V-FLUTE split: 4,578 training, 726 validation, and 723 test samples. The rationale memory is built from the FLUTE textual training set only. To rule out data leakage, we conduct overlap and retrieval contamination audits between the FLUTE memory and the V-FLUTE test set; exact overlaps are extremely rare and no retrieved top- K rationale exactly matches any V-FLUTE test explanation. Detailed results are provided in the Appendix.

4.2 Baselines

We compare RIFT 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 Explanation-Score falls below the respective threshold, even when the predicted label is correct. Explanation-Score 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 Settings

We use LLaVA-v1.6-Vicuna-7B as the backbone. For feature extraction, a frozen CLIP ViT-L/14 (Radford et al., 2021) is used to obtain 768-dimensional image patch features, and RoBERTa-large (Liu et al., 2019) is used to encode the description and claim into 1024-dimensional textual representations. For the external relation transfer step, all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) is used to encode the FLUTE relation keys and online queries into 384-dimensional dense vectors. We apply LoRA to the LLaVA language model for parameter-efficient fine-tuning.

Training is conducted for 4 epochs with a maximum sequence length of 2048, a projector learning rate of $3e-5$, a language model learning rate of $2e-4$, and a warmup ratio of 0.03. The latent prefix contains 4 internal relation tokens and 2 external relation tokens, and the top 8 samples are used in the top- K external rationale retrieval process. All the experiments are conducted on a single NVIDIA A40 GPU with 48 GB of memory.

4.5 Main Results

Table 1 presents the main experimental results. RIFT achieves the best performance across all three metrics, with F1@0, F1@53, and F1@60 values of 88.52, 76.07, and 59.47, respectively. Compared with those of MAPPER, the F1@0, F1@53, and F1@60 scores of RIFT are increased by 3.98 ($p = 0.0015$), 6.38 ($p = 0.0002$), and 7.75 ($p < 0.0001$) points, respectively.¹ Moreover, RIFT substantially outperforms the strongest fine-tuned baseline, LLaVA-1.5-7B-eViL+VF, with gains of 13.61, 13.73, and 10.67 points on these three metrics, respectively. These results demonstrate that combining internal multimodal relation modeling with external figurative relation transfer effectively improves the visual figurative entailment performance.

¹ p is computed using a two-sided paired randomization test with 10,000 permutations over the same 723 test instances.

Model	F1@0	F1@53	F1@60
RIFT	88.52	76.07	59.47
w/o Routing Gate	87.56	75.52	56.43
w/o External Transfer	87.21	75.78	57.11
w/o Internal Modeling	86.82	73.99	55.46
w/o Both (LLaVA + CD)	85.10	71.36	54.05
LLaVA SFT	80.91	67.90	49.27

Table 2: Ablation results on V-FLUTE. Routing Gate denotes the confidence routing gate. CD denotes a claim-conditioned description.

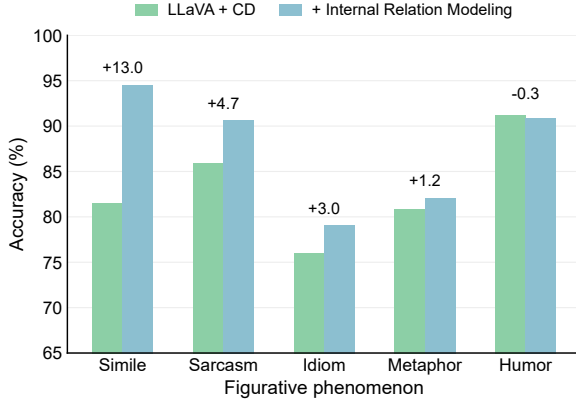


Figure 3: Per-category accuracy comparison between LLaVA + CD and LLaVA + CD + Internal Relation Modeling on the V-FLUTE test set.

Notably, the gains are more pronounced for the F1@53 and F1@60 metrics, indicating that the advantages of RIFT extend beyond label prediction to explanation-constrained evaluation. This suggests that the internal relation modeling component helps the model better capture relational cues among the image, description, and claim, whereas the external figurative relation transfer component provides reusable nonliteral reasoning patterns. Together, these two components increase the consistency between the entailment prediction and the generated explanations.

4.6 Ablation Study

Table 2 presents the results of the ablation experiments. The w/o Both (LLaVA + CD) variant outperforms the LLaVA SFT across all three metrics, showing that the claim-conditioned descriptions provide useful claim-relevant visual cues for figurative image-claim understanding. The model in which the external figurative relation transfer component is removed while retaining the internal relation modeling component outperforms the w/o Both model, indicating that the relation-gated MoE

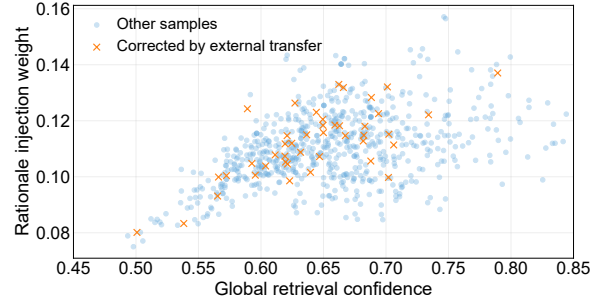


Figure 4: Scatter plot of the rationale injection weights versus the global retrieval confidence scores for V-FLUTE test samples.

strategy effectively captures multimodal relations among the image, description, and claim. Similarly, the model in which the external figurative relation transfer component is retained while the internal relation modeling component is removed outperforms the w/o Both model, suggesting that the FLUTE-based rationale retrieval component introduces transferable nonliteral reasoning patterns. The full RIFT model achieves the best performance in terms of all the metrics, confirming the complementarity between the internal relation modeling and external figurative relation transfer components. Finally, the performance of the model in which the confidence routing gate is removed is notably decreased, especially in terms of the F1@60 metric, indicating that the adaptive routing strategy helps suppress low-relevance retrievals while improving the explanation-constrained metrics.

4.7 Effect of the Internal Relation Modeling Strategy across Figurative Phenomena

We next analyze the effect of the internal relation modeling component across different figurative phenomena (Figure 3). Compared with the LLaVA + CD baseline, the addition of this module yields clear improvements in categories that require stronger cross-modal nonliteral alignment, such as similes, sarcasm, and idioms. For example, the accuracy on the simile category increases from 81.48% to 94.44%. This suggests that the internal relation modeling component helps the model capture relational structures among the image, description, and claim rather than simply recognizing salient visual objects. For the simile and idiom categories, the model must map concrete visual actions or scenes to abstract comparative or idiomatic meanings. For sarcasm, semantic tension between the visual evidence and the claim needs

Input	MAPPER	RIFT
	Dataset: IRFL (idiom) Claim: "mend fences" Gold label: Entailment Visual description: The image shows two hands shaking. The hand on the left is black and the hand on the right is white. They are both holding each other's hands. The background is a white wall with a textured surface.	
Predicted label	Contradiction	Entailment
ExplanationScore	65.07	74.71
Model output	contradiction. The image depicts two people shaking hands, which is a gesture typically used to show agreement, respect, or the beginning of a relationship . The idiom "mend fences" suggests making amends or reconciling with someone, which contradicts the image because the people are not in a state of conflict or separation that would require them to 'mend fences.' Instead, they appear to be coming to an agreement or starting a collaboration .	entailment. The image depicts two people shaking hands, which is often seen as a gesture of agreement or reconciliation . The idiom "mend fences" suggests making amends or repairing relationships , which is entailed by the image as the act of shaking hands can symbolize coming to an understanding or resolving differences , thus implying the fences have been mended .
Ground explanation	The image depicts two individuals shaking hands, a common gesture of reconciliation, agreement, or unity. The idiom "mend fences" signifies repairing damaged relationships or resolving disagreements. The act of shaking hands in the image entails the concept of mending fences as it visually represents the idea of making peace or rebuilding connections after a conflict.	

Figure 5: A case study from the IRFL idiom subset.

to be identified, while abstract attribute mapping is often required for metaphors. Compared with the LLaVA + CD baseline, the relation-gated multimodal MoE strategy better integrates multimodal evidence at the sample level and captures nonliteral semantic correspondences. The smaller gain on the humor category suggests that this component is more advantageous on categories requiring stronger relational reasoning.

4.8 Effect of External Relation Transfer on Error Correction and Rationale Routing

As shown in Figure 4, the injection weight of the external rationale varies across samples, ranging from approximately 0.075 to 0.157. This finding indicates that the confidence routing gate performs sample-level adjustment rather than uniformly injecting FLUTE rationales. The injection weight generally increases as the global retrieval confidence increases, although the relationship is not strictly linear. The samples corrected by the external relation transfer module are concentrated mainly in regions with moderate to high retrieval confidence and relatively high injection weights, suggesting that external rationales are most helpful when their relation patterns match the current image-claim pair.

These results show that the external relation transfer module can selectively use figurative knowledge from FLUTE. Specifically, reliable retrieval signals receive stronger external guidance, whereas weak retrieval signals are assigned lower injection weights. With this approach, the model benefits from transferable nonliteral reasoning pat-

terns while reducing noise from low-relevance rationales.

4.9 Case Study

A representative example from the IRFL idiom subset is shown in Figure 5. The image depicts two people shaking hands, and the claim contains the idiom "mend fences," which means repairing a relationship or resolving a conflict. MAPPER predicts this sample as "contradiction," likely because the image does not explicitly show a prior conflict or disagreement. Although MAPPER recognizes the handshake action, its reasoning remains largely literal, with the absence of explicit conflict cues considered evidence against the idiomatic meaning.

In contrast, RIFT correctly predicts "entailment" by recognizing that a handshake can symbolize reconciliation and the repair of relationships. Rather than requiring explicit visual evidence of conflict, RIFT maps the concrete social action to the abstract relational meaning of "mend fences". This suggests that the external figurative relation transfer module helps the model move beyond the interpretation of literal visual actions toward nonliteral semantic entailment, whereas the internal relation modeling component keeps the inference grounded in the current visual evidence.

5 Conclusion

In this paper, visual figurative entailment is addressed via internal relation modeling and external figurative relation transfer strategies. We propose RIFT, an internal-external relation reasoning frame-

work that models relations among the image, claim-conditioned visual description, and textual claim via a relation-gated multimodal MoE strategy while transferring reusable relation-level reasoning patterns from FLUTE through a rationale retrieval process. A confidence routing gate further controls the injection strength of the external rationales on the basis of the retrieval reliability. The results of experiments on the V-FLUTE dataset show that RIFT outperforms existing baselines, with particularly strong gains on explanation-constrained metrics. The results of the ablation studies confirm the complementarity of the internal relation modeling and external figurative relation transfer components. In future work, more fine-grained relation memory construction and retrieval strategies for different figurative phenomena will be explored.

Limitations

Although RIFT achieves strong performance on visual figurative entailment tasks, several limitations should be considered. First, the external rationale retrieval module uses a lightweight MiniLM encoder for efficiency, which may limit the fine-grained relation matching performance compared with the use of stronger retrievers or rerankers. Second, RIFT relies on claim-conditioned descriptions generated by a frozen multimodal model. While these descriptions are useful for revealing claim-relevant visual cues, they may also introduce hallucinated or claim-biased information. Third, RIFT is currently built upon a single multimodal backbone, and its effectiveness across different backbone architectures and model scales remains unexplored. Fourth, the confidence routing gate is optimized through downstream supervision rather than explicit retrieval quality labels; thus, its calibration and stability must be further investigated.

Future work should focus on the exploration of broader figurative knowledge sources, the use of stronger retrieval architectures, and more detailed analyses of the efficiency and reliability of the proposed approach.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62562057), the Key Research and Development Program of Xinjiang Autonomous Region (Grant No. 2024B03041), and the Xinjiang Key Laboratory of Intelligent Computing and Smart Appli-

cations. We also thank the Computing and Data Center of Xinjiang University for providing the computational resources used in the extensive experiments reported in this study.

References

- Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. Flute: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159.
- Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388.
- Jiale Chen, Qihao Yang, Xuelian Dong, Xiaoling Mao, and Tianyong Hao. 2024. A textual modal supplement framework for understanding multi-modal figurative language. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 85–91.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tjong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.
- Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.
- Virginie Do, Oana-Maria Camburu, Zeynep Akata, and Thomas Lukasiewicz. 2020. e-snli-ve: Corrected visual-textual entailment with natural language explanations. *arXiv preprint arXiv:2004.03744*.
- J Hessel, A Marasovic, JD Hwang, L Lee, J Da, R Zellers, R Mankoff, and Y Choi. 2022. Do androids laugh at electric sheep. *Humor Understanding Benchmarks from The New Yorker Caption Contest*. *arXiv*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

- EunJeong Hwang and Vered Shwartz. 2023. Meme-cap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445.
- Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. 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*, pages 1244–1254.
- Shreyas Kulkarni, Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2024. A report on the figlang 2024 shared task on multimodal figurative language. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 115–119.
- George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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)*, pages 4582–4597.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.
- Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. 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)*, pages 1–23.
- Ekaterina V Shutova. 2011. Computational approaches to figurative language. Technical report, University of Cambridge, Computer Laboratory.
- Yuan Tian, Minzheng Wang, Nan Xu, and Wenji Mao. 2025. Imara: An imaginative frame augmented method for low-resource multimodal metaphor detection and explanation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3953–3967.
- Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3296–3304.
- Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. 2025. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. *arXiv preprint arXiv:2506.07218*.
- Qihao Yang and Xuelin Wang. 2024. Figclip: A generative multimodal model with bidirectional cross-attention for understanding figurative language via visual entailment. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 92–98.
- Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13040–13051.
- Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. Irfl: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058.
- Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multi-modal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Type	Dataset	Train	Val	Test
Metaphor/Simile	HAIVMet(Chakrabarty et al., 2023)	649	107	101
Metaphor/Simile	IRFL(Yosef et al., 2023)	912	117	120
Idiom	IRFL(Yosef et al., 2023)	170	100	100
Sarcasm	MuSE(Desai et al., 2022)	830	106	106
Humor	MemeCap(Hwang and Shwartz, 2023)	1566	196	196
Humor	NYCartoons(Hessel et al., 2022)	451	100	100
Total	-	4578	726	723

Table 3: Data distribution of V-FLUTE across figurative types and source datasets.

Check item	Count	Ratio
V-FLUTE test samples	723	100%
Exact claim match	1	0.14%
Exact gold explanation match	0	0.00%
Exact relation text match	0	0.00%
Nearest neighbor cosine > 0.90	3	0.41%
Retrieved top- K rationales	5,784	-
Exact retrieved explanation match	0	0.00%
Retrieved explanation ROUGE-L > 0.80	0	0.00%

Table 4: Data leakage and retrieval contamination check between the V-FLUTE test set and the FLUTE rationale bank. Exact match checks compare V-FLUTE test claims, gold explanations, and relation texts with FLUTE hypotheses, explanations, and relation texts after normalization. Retrieved top- K rationales refer to the actual rationales retrieved by the model during inference.

A Data Distribution of V-FLUTE

Table 3 reports the data distribution of V-FLUTE across figurative phenomena and source datasets, covering metaphor/simile, idiom, sarcasm, and humor. The distribution is imbalanced across subsets: MemeCap contributes the largest number of training samples, whereas IRFL idioms contain only 170 training instances. This reflects the low resource nature of visual figurative entailment, where certain figurative phenomena lack sufficient image text supervision. Such data scarcity motivates the incorporation of external rationale retrieval to provide complementary relation level reasoning signals beyond what the training data alone can supply.

B Data Leakage and Retrieval Contamination Check

Because RIFT retrieves external textual rationales from the FLUTE relation bank during the inference stage on V-FLUTE, we examine whether the observed gains could be attributed to duplicated or near-duplicated information from the test set. We conduct two audits. First, after lowercasing the text, removing punctuation, and merging redundant whitespace, we compare the V-FLUTE test claims, reference explanations, and relation text with the FLUTE hypotheses, explanations, and relation text. Second, we compute nearest-neighbor similarities

between the V-FLUTE test queries and FLUTE entries in the retrieval embedding space. We also investigate the actual top- K rationales retrieved by the full model, as contamination of the retrieved rationales would have the most direct impact on the evaluation results.

Table 4 shows the results. Exact overlaps are extremely rare: among 723 V-FLUTE test samples, only one claim exactly matches a FLUTE hypothesis, and there are no exact matches for the reference explanations or relation texts. In the semantic nearest-neighbor analysis, only three test samples have a maximum cosine similarity above 0.90 with any FLUTE entry. More importantly, among all 5,784 top- K rationales that are actually retrieved by the model, none exactly match a V-FLUTE reference explanation, and none has an explanation ROUGE-L score above 0.80. These results indicate that the external relation transfer strategy does not access duplicate test explanations, suggesting that the performance gains of RIFT are unlikely to result from data leakage.

C Prompt Templates

We use two natural language prompts in our V-FLUTE experiments. The first prompt is used to generate claim-conditioned visual descriptions. The second prompt is used as the final instruction prompt for supervised fine tuning and inference.

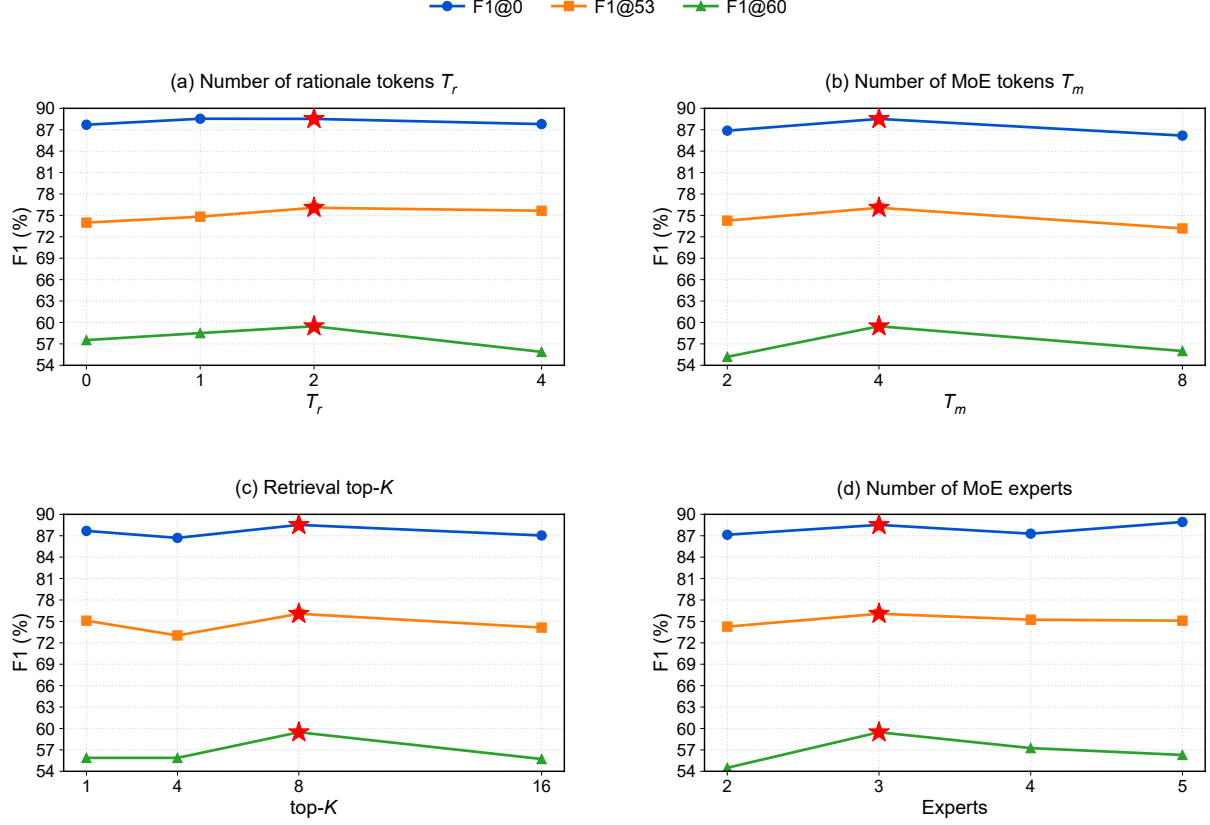


Figure 6: Parameter sensitivity analysis on the V-FLUTE test set.

Phenomenon	Dataset	Human Avg.	LLaVA eViL+VF	RIFT	Δ Human
Metaphor/Simile	HAIVMET	78.84	81.25	83.97	+5.13
Metaphor/Simile	IRFL (metaphor/simile)	94.36	77.78	90.80	-3.56
Idioms	IRFL (idiom)	89.26	49.74	72.54	-16.72
Sarcasm	MuSE	68.89	85.42	92.45	+23.56
Humor	MemeCap	100.00	78.03	89.78	-10.22
Humor	NYCartoons	71.43	47.83	100.00	+28.57
Overall	All	89.09	77.26	88.52	-0.57

Table 5: Subset level F1@0 results on V-FLUTE. Human Avg. and LLaVA eViL+VF are taken from the original V-FLUTE evaluation table. Δ Human denotes the difference between RIFT and the reported human average.

During supervised fine tuning, the target response consists of the gold label followed by the gold explanation. During inference, only the input prompt is provided to the model.

Description generation prompt. For each image-claim pair, we use LLaVA-v1.6-Vicuna-7B as a visual interpreter to generate a claim-conditioned visual description. The prompt used for description generation is:

Here is an image accompanied by the claim: '{claim}'. Please describe what you see in the provided picture in detail. Mention any visual elements that relate to this claim.

The generated text is used as the *description* field in the final V-FLUTE instruction prompt.

Final training and inference prompt. The final prompt takes the image, the generated visual description, and the textual claim as input. It asks the model to predict whether the claim is entailed or contradicted by the image and to generate the corresponding explanation:

<image>

The description and visual analysis of this picture is '{description}'. The claim of this picture is '{claim}'. You need to predict whether the claim of this picture is 'entailment' or 'contradiction' firstly according to the picture and its description. Then you need to give an explanation for the prediction. The prediction and the explanation are related to the meaning of the figurative language expression.

Dataset	LLaVA SFT	LLaVA + CD	+ Internal	+ External	RIFT
HAIVMET	79.47	82.93	83.11	85.05	83.97
IRFL (metaphor/simile)	75.69	78.92	86.57	88.28	90.80
IRFL (idiom)	44.79	75.52	78.90	71.44	72.54
MuSE	88.68	85.79	90.56	89.58	92.45
MemeCap	85.20	86.73	86.16	86.71	89.78
NYCartoons	48.98	100.00	100.00	100.00	100.00
Overall	80.91	85.10	87.21	86.82	88.52

Table 6: Subset level ablation results measured by F1@0. LLaVA + CD denotes LLaVA with claim conditioned descriptions. + Internal adds internal relation modeling with the relation gated multimodal MoE on top of LLaVA + CD. + External adds external FLUTE rationale transfer on top of LLaVA + CD without the internal MoE module. RIFT combines both internal relation modeling and external figurative relation transfer.

Your response must follow the format shown below:

{Prediction}. {Explanation}.

For supervised fine tuning, the target response is:

{gold label}. {gold explanation}.

where the gold label is either *entailment* or *contradiction*. The same input prompt is used during inference. For evaluation, we parse the first sentence of the generated response as the predicted label and treat the remaining text as the explanation.

D Parameter Sensitivity Analysis

To analyze the influence of key parameters on RIFT, we examine the number of rationale tokens T_r , the number of MoE tokens T_m , retrieval top- K , and the number of MoE experts, as shown in Figure 6. As shown in Figure 6(a), when T_r increases from 0 to 2, the model improves overall on all three metrics, suggesting that a suitable number of rationale tokens helps supplement intermediate reasoning information. However, when T_r further increases to 4, F1@60 drops noticeably, indicating that excessive reasoning tokens may introduce redundant noise. As shown in Figure 6(b), the model performs best when $T_m = 4$, suggesting that a moderate number of MoE tokens provides sufficient relation representation capacity, whereas too few or too many tokens can limit key information modeling. As shown in Figure 6(c), retrieval with $K = 8$ achieves the best or near-best performance, indicating that an appropriate number of external examples can effectively supplement figurative relation knowledge. When K is too small or too large, the model faces insufficient knowledge coverage or increased retrieval noise, respectively. As shown in Figure 6(d), using 3 MoE experts performs best

on F1@53 and F1@60, suggesting that this setting better distinguishes different multimodal figurative relation patterns. Although further increasing the number of experts may slightly improve F1@0, it reduces overall performance under explanation constraints.

Overall, RIFT achieves the most stable comprehensive performance with $T_r = 2$, $T_m = 4$, $K = 8$, and 3 MoE experts, striking a balance among reasoning information, expert capacity, external knowledge coverage, and noise control.

E Subset Results

We further report subset level results on V-FLUTE to analyze the performance of RIFT across different figurative phenomena and source datasets. Table 5 compares RIFT with the reported human average and the LLaVA eViL+VF baseline from the original V-FLUTE evaluation table. Note that the LLaVA eViL+VF result reported in Table 5 is evaluated on the 99-example human-evaluation subset used in the original V-FLUTE paper, whereas Table 1 reports results on the full V-FLUTE test set containing 723 examples. We report F1@0 for this analysis.

Compared with LLaVA eViL+VF, RIFT improves on all subsets, showing that the proposed internal external relation reasoning framework generalizes across different figurative phenomena and source datasets. The largest gains over LLaVA eViL+VF appear on NYCartoons, IRFL idioms, IRFL metaphor/simile, and MemeCap, suggesting that RIFT can provide substantial benefits for cases requiring non literal interpretation beyond local visual or lexical matching. Compared with the reported human average, RIFT achieves a close overall score, reaching 88.52 F1@0 versus 89.09 for humans. However, the gap remains noticeable on IRFL idioms and MemeCap, where RIFT is 16.72 and 10.22 points below the human average, respec-

Phenomenon	Dataset	N	Human Avg. (0–100)	95% Bootstrap CI
Metaphor/Simile	HAIVMET	14	66.96	[49.11, 82.14]
Metaphor/Simile	IRFL (metaphor/simile)	13	90.38	[74.04, 100.00]
Idioms	IRFL (idiom)	17	63.97	[41.91, 84.56]
Sarcasm	MuSE	14	71.43	[55.36, 84.82]
Humor	MemeCap	27	68.98	[56.48, 80.09]
Humor	NYCartoons	14	70.54	[58.04, 81.25]
Overall	All	99	71.21	[64.52, 77.53]

Table 7: Human evaluation of RIFT-generated explanations across figurative phenomena and source datasets. Human Avg. denotes the mean score assigned by the two co-authors after linear rescaling to a 0–100 scale. Confidence intervals were estimated using 20,000 sample-level bootstrap resamples.

tively. This indicates that idiomatic reasoning and some meme style humor remain challenging for the current model.

Table 6 provides a subset level view of different model variants. Adding claim conditioned descriptions improves the LLaVA SFT on most subsets, especially on IRFL idioms. Based on LLaVA + CD, adding internal relation modeling further improves metaphor/simile, idiom, and sarcasm performance, while adding external FLUTE rationale transfer is particularly helpful on HAIVMET and IRFL metaphor/simile. The full RIFT model achieves the best overall performance and obtains the best results on IRFL metaphor/simile, MuSE, MemeCap, and NYCartoons, further supporting the complementarity between internal multimodal relation modeling and external figurative relation transfer.

A closer look at IRFL idioms reveals that this subset remains challenging. The + External variant obtains 71.44 F1@0, which is lower than both LLaVA + CD and + Internal, and the full RIFT model also remains below LLaVA + CD on this subset. This suggests that external FLUTE rationale transfer may be less reliable for idiom samples, where retrieved textual rationales do not always align with the specific visual action or scene. Since idioms often depend on highly conventionalized and context dependent meanings, transferring relation patterns from textual data to visual cases remains difficult. Although RIFT achieves the best overall performance across all subsets, its 16.72 point gap from the reported human average on IRFL idioms indicates that idiomatic visual figurative entailment requires more fine grained retrieval and grounding mechanisms.

F Human Evaluation

To complement the automatic evaluation metrics, we conducted a human evaluation of the expla-

nations generated by RIFT. We selected 99 test instances using stratified sampling across the six phenomenon–dataset groups listed in Table 7. This sampling strategy ensured coverage of metaphor and simile, idiom, sarcasm, and humor. Two co-authors independently evaluated all selected instances. Each evaluator has academic training in linguistics and holds at least a master’s degree.

Evaluation Protocol. The co-authors rated each explanation on a five-point scale according to four criteria: (1) the correctness of the predicted entailment label, (2) the interpretation of the intended figurative meaning, (3) the extent to which the reasoning was supported by the image and textual evidence, and (4) semantic consistency with the reference explanation. A score of 1 indicated an incorrect, irrelevant, or unsupported explanation; 2 indicated a mostly incorrect explanation with a major reasoning error; 3 indicated a partially correct but incomplete explanation; 4 indicated a mostly correct explanation with only minor omissions or imprecision; and 5 indicated a complete, coherent, and well-supported explanation. Scores of 4–5 were treated as adequate and scores of 1–3 as inadequate. When the predicted label contradicted the gold label, the score was generally limited to at most 2 unless the reference annotation was clearly problematic.

Score Aggregation and Agreement. Let $s_{ij} \in \{1, \dots, 5\}$ denote the score assigned to sample i by co-author j . We first averaged the two ratings and then linearly rescaled the resulting score to a 0–100 scale:

$$H_i = \frac{\left(\frac{1}{2} \sum_{j=1}^2 s_{ij}\right) - 1}{4} \times 100 \quad (10)$$

We report the mean of H_i for each stratum and for the complete evaluation set. The corresponding 95% confidence intervals were estimated using

Model	Accuracy	Δ Acc.	95% CI	Macro-F1	Δ F1	95% CI / p -value
Vanilla LLaVA	57.68	–	–	46.05	–	–
RIFT	87.23	+29.55	[28.38, 30.72]	87.22	+41.17	[40.08, 42.25] / 0.0001

Table 8: Zero-shot cross-dataset evaluation on the binary entailment/contradiction subset of e-SNLI-VE. The improvements are computed relative to Vanilla LLaVA. The accuracy improvement is also significant under paired randomization testing ($p = 0.0001$).

20,000 sample-level bootstrap resamples. For the binary adequacy judgments, the two co-authors achieved an observed agreement of 84.85%, corresponding to Cohen’s $\kappa = 0.636$ with a 95% bootstrap confidence interval of [0.465, 0.792]. This indicates substantial inter-annotator agreement.

As shown in Table 7, RIFT obtains an overall human evaluation score of 71.21, with a 95% bootstrap confidence interval of [64.52, 77.53]. The highest average score is observed on the IRFL metaphor/simile subset (90.38), while the scores on the remaining subsets range from 63.97 to 71.43. Given the relatively small number of samples in each stratum, the subset-level results should be interpreted together with their confidence intervals.

G Cross-Dataset Generalization

We further evaluate the cross-dataset generalization of RIFT on e-SNLI-VE, an independent benchmark for visual entailment with natural-language explanations. Unlike V-FLUTE, which focuses on figurative visual entailment, e-SNLI-VE primarily evaluates general visual entailment, thereby introducing both dataset and task-domain shifts. To align its original three-class label space with the binary setting of V-FLUTE, we remove the neutral instances from the official test split. This results in 10,939 test instances, comprising 5,218 entailment and 5,721 contradiction examples over 1,000 unique images. We directly evaluate the V-FLUTE-trained RIFT checkpoint on this subset without target-domain training, fine-tuning, hyperparameter selection, or access to the e-SNLI-VE training set.

As shown in Table 8, RIFT achieves an accuracy of 87.23 and a Macro-F1 score of 87.22, compared with 57.68 and 46.05, respectively, for the original LLaVA backbone. We assess the differences using 10,000 paired bootstrap resamples for 95% confidence intervals and paired randomization tests for statistical significance. RIFT improves accuracy by 29.55 points (95% CI: [28.38, 30.72], $p = 0.0001$) and Macro-F1 by 41.17 points (95%

CI: [40.08, 42.25], $p = 0.0001$). These results provide evidence that the task-specific reasoning capability learned on V-FLUTE generalizes to an independent visual-entailment dataset.