

Enhancing Pun Meme Understanding with Visual Semantic Cues

Meng Li, Zhongqing Wang*, Guodong Zhou

Natural Language Processing Lab, Soochow University, Suzhou, China

lm_9841@163.com, wangzq@suda.edu.cn, gdzhou@suda.edu.cn

Abstract

Internet memes have become important carriers of online expression, and pun memes, as a typical multimodal category, pose significant challenges to machine understanding due to their reliance on textual ambiguity and cross-modal interaction with visual content. Existing multimodal models for pun meme understanding suffer from classification bias, weak sentiment perception, and performance gaps compared to human comprehension, mainly due to neglecting fine-grained correlation cues for semantic reasoning. To address this, we propose a fusion framework enhanced by multi-source auxiliary cues, including visual salient perception, textual deep semantics, and abnormal pun linguistic features, which are embedded into the multimodal reasoning pipeline. Experiments show that our framework stably optimizes vision-language models, improving accuracy in pun meme detection and implicit sentiment fine-grained analysis.

1 Introduction

Internet memes have evolved into core mainstream carriers of online interactive expression, integrating visual content and textual captions to implicitly deliver situational humor, personal attitudes, emotional tendencies and targeted social opinions (Bauckhage, 2011; Laineste and Voolaid, 2016; Denisova, 2019).

As a typical high-correlation category of multimodal memes, pun memes (Šimon, 2019; Xu et al., 2025) occupy a pivotal core position in relevant research tasks, and they pose prominent challenges to multimodal understanding models. The core semantic connotation of pun memes is mostly constructed through textual ambiguity and cross-modal interaction with visual content, which creates inherent barriers to accurate machine recognition (Chung et al., 2024). Furthermore, meme

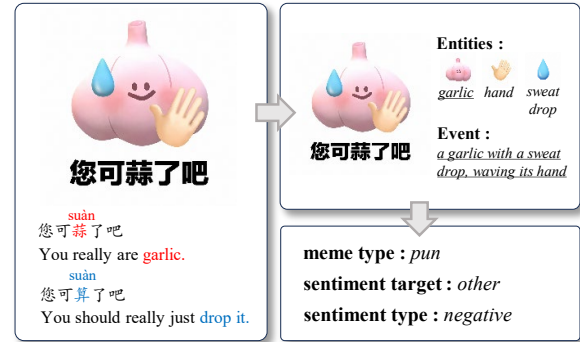


Figure 1: An example of visual semantic cues for pun meme understanding.

textual captions usually serve as the linguistic trigger of the pun interpretation, while the matched visual scene provides contextual support for grounding and calibrating the intended meaning (Vásquez and Aslan, 2021; Kiela et al., 2020). As shown in Fig. 1, the caption “您可蒜了吧” replaces “算” in the expression “您可算了吧” with the homophonic character “蒜”, both pronounced *suàn*. The substituted caption literally means “you really are garlic”, which is directly grounded in the anthropomorphic garlic shown in the image. In contrast, the intended expression “您可算了吧” means “you should really just drop it”, conveying a dismissive and mildly sarcastic attitude. The sweat drop and hand-waving gesture further provide an event-level cue for this negative emotional undertone. This example further demonstrates that understanding pun memes requires grounding textual ambiguity in visual semantic cues.

Prior research on pun meme understanding has made preliminary progress in constructing multimodal benchmark settings and baseline parsing frameworks. However, existing multimodal models still suffer from classification bias, weak perception of implicit sentiment behind pun expressions, and a considerable gap from human cross-modal comprehension (Xu et al., 2025). The essential

*Corresponding author.

limitation of current research lies in the neglect of multiple key fine-grained correlation cues required for pun meme semantic reasoning (Wang et al., 2024; Zheng et al., 2025). On the visual side, meme images contain salient entities, behavioral states, and affective visual signals that guide the interpretation of ambiguous captions. On the textual side, pun-driven captions involve layered semantic structures and subtle affective implications, which are difficult to align with visual context through conventional multimodal fusion strategies.

To remedy the above research defects and fill the technical gap, this paper proposes a multi-source auxiliary cue enhanced framework for targeted pun meme understanding. Different from conventional end-to-end brute-force reasoning modes based on original image-text pairs, the proposed framework explicitly introduces structured auxiliary cues closely related to pun construction, and embeds them into the multimodal reasoning process as complementary evidence. This design enables the model to perform more targeted, interpretable, and accurate reasoning over complex cross-modal ambiguous semantics in pun memes.

Comparative experimental validations on the authoritative PUNMEMECN benchmark dataset (Xu et al., 2025) verify that the structured multi-source auxiliary cue embedding strategy can steadily optimize the overall learning performance of mainstream vision-language models (VLMs), and achieve consistent accuracy improvement in two core downstream tasks, including pun meme intelligent detection and implicit sentiment fine-grained analysis.

2 Related Works

2.1 Pun Understanding

Pun understanding aims to identify and interpret expressions that exploit lexical ambiguity, polysemy, or phonological similarity to convey multiple meanings. Early computational studies were primarily text-based, relying on lexical-semantic modeling, word-sense disambiguation, or sequence modeling to analyze pun ambiguity (Miller and Gurevych, 2015; Miller and Turković, 2016; Kao et al., 2016; Zou and Lu, 2019). Subsequent neural and large language model based research further extended this line of work to pun detection, interpretation, explanation, and generation (Yu et al., 2018; Sun et al., 2022; Tian et al., 2022; Xu et al., 2024; Zangari et al., 2025).

More recently, multimodal studies have begun to investigate the role of visual information in pun interpretation. Existing research has explored visual grounding for textual pun disambiguation, multimodal pun reconstruction, and textual-visual pun translation (Chung et al., 2024; Ma et al., 2025; Zhang et al., 2025). These studies establish that visual context can help distinguish competing meanings of ambiguous expressions. Nevertheless, their primary concern lies in resolving or transforming textual puns with visual assistance, rather than examining how fine-grained visual semantics support the interpretation of naturally occurring pun memes.

2.2 Meme Detection and Analysis

Internet memes combine visual content and textual captions to convey humor, attitudes, emotions, and social opinions (Denisova, 2019; Hakoköngäs et al., 2020). Accordingly, multimodal meme analysis has been explored in tasks such as hateful meme detection, offensive content identification, sentiment analysis, humor recognition, metaphor understanding, and fine-grained semantic interpretation (Suryawanshi et al., 2020; Sharma et al., 2020; Hossain et al., 2022; Hee et al., 2023; Wang et al., 2024). However, although existing multimodal methods have advanced general meme analysis, they are not specifically designed for pun memes, where ambiguous captions usually function as linguistic triggers and specific visual cues provide essential support for interpretation (Xu et al., 2025). As a result, holistic image-text modeling may fail to capture the fine-grained correspondence between pun expressions and key visual cues.

Motivated by this distinction, our work shifts the focus from general multimodal fusion to the semantic correspondence between pun-triggering captions and their relevant visual content. Specifically, we examine how visual objects and event-level states relate pun expressions to their intended meanings and affective implications. This perspective extends prior benchmark-oriented and end-to-end multimodal studies by moving from whether visual context is useful to which visual semantics support pun interpretation and sentiment judgment.

3 Task Formulation

As illustrated in Fig. 3, given a meme image I and its caption C , we formulate pun meme understanding as a joint task with input $x =$

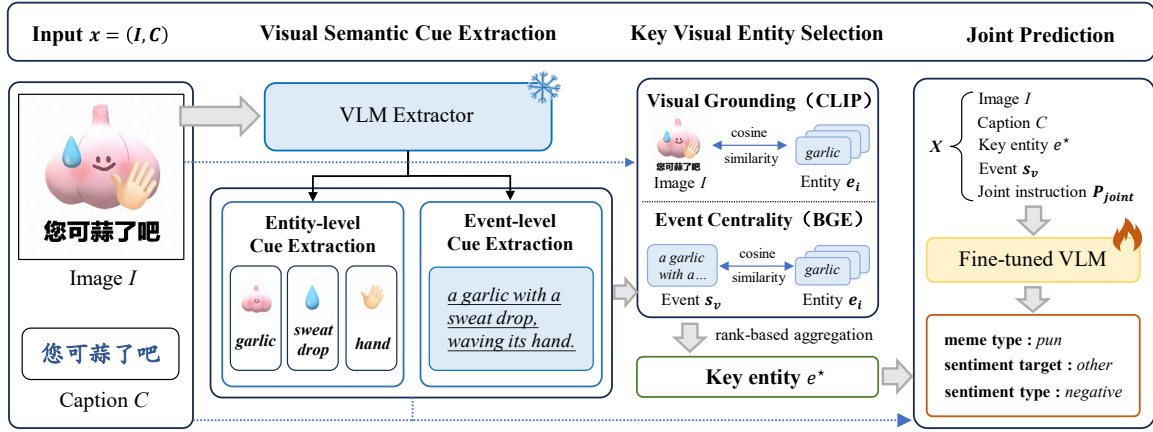


Figure 2: Overview of the proposed model.

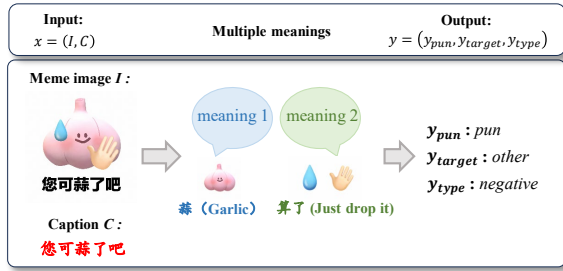


Figure 3: Task formulation of pun meme understanding.

(I, C) and output $y = (y_{\text{pun}}, y_{\text{target}}, y_{\text{type}})$. Here, $y_{\text{pun}} \in \{\text{pun}, \text{non-pun}\}$ denotes pun meme detection. For pun memes, $y_{\text{target}} \in \{\text{self}, \text{other}, \text{both}\}$ and $y_{\text{type}} \in \{\text{positive}, \text{neutral}, \text{negative}\}$ denote the sentiment target and type, respectively.

A meme is regarded as a *pun meme* when its caption evokes multiple meanings through homophony or polysemy, with at least one interpretation grounded in the accompanying image; otherwise, it is categorized as a *non-pun meme*. For instance, the example in Fig. 3, “您可蒜了吧”, uses the homophony between “蒜” and “算” (*suàn*), where the depicted garlic grounds the literal reading and the intended expression conveys a negative attitude toward an external target.

Figure 4 provides additional examples to make this interaction more intuitive. In “放青松/放轻松” (*fàng qīngsōng*), a hand putting down a pine tree grounds the literal reading of “青松” (*pine tree*), while its homophone “轻松” conveys the intended meaning of *relax*. Likewise, in “举豹了/举报了” (*jǔ bào le*), a person holding up a leopard supports the literal reading, while the homophonic expression means reporting someone. “真菜” (*zhēn cài*) illustrates polysemy: a magnifying glass examining

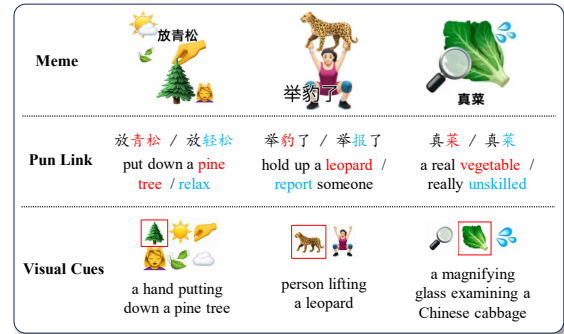


Figure 4: Illustrative examples of pun memes with visual cues.

a Chinese cabbage reinforces its literal reading as *really a vegetable*, whereas its colloquial meaning conveys a negative evaluation of being unskilled. These examples highlight that visual content is not merely decorative, but provides key evidence for linking a literal reading to the intended pun meaning.

Although these examples are intuitive to human readers, interpreting such visually grounded ambiguity remains challenging for multimodal models. First, the model must distinguish deliberate word-play from ordinary expressions. Second, it must identify the visual elements relevant to the ambiguous caption. Third, sentiment prediction depends on the resolved pun meaning and its associated visual states. Therefore, pun meme understanding requires jointly modeling caption ambiguity, task-relevant visual semantics, and affective implications.

4 Method

As shown in Fig. 2, we propose a visual semantic cue enhancement framework for pun meme understanding. Rather than relying solely on global image–caption interaction, our framework identifies visual evidence that grounds caption-induced ambiguity, thereby enabling visual-cue-guided reasoning. Specifically, given an image I and its caption C , our framework constructs structured visual evidence from two complementary perspectives: candidate entities provide possible visual referents of the pun, while an event-level cue captures the contextual semantics surrounding them. It then selects a key visual entity by considering both its visual grounding in the image and its semantic centrality to the depicted event. Finally, the selected entity and event-level cue are introduced into vision-language reasoning to focus the model on pun-relevant visual semantics and support joint pun detection and sentiment analysis.

4.1 Visual Semantic Cue Extraction

Pun meme understanding requires visual evidence that supports the interpretation of ambiguous captions. To explicitly represent such evidence, we extract two types of visual semantic cues from the meme image: entity-level cues and event-level cues. Entity-level cues describe visible objects, characters, or other semantically meaningful elements, while event-level cues capture the primary action, state, or contextual relation involving these elements. These structured cues are subsequently used for key entity selection and cue-conditioned reasoning.

Entity-level Cue Extraction. Given a meme image I , we use a vision-language model as the visual cue extractor to obtain a candidate entity set:

$$E = f_{\text{VLM}}(I, P_{\text{ent}}) = \{e_1, e_2, \dots, e_m\}, \quad (1)$$

where E denotes the candidate entity set, and P_{ent} denotes the prompt. Instead of directly outputting a single entity, we retain multiple candidates at this stage, which increases the chance of covering the truly relevant entity and reduces the risk of early selection errors.

Event-level Cue Extraction. In parallel, we perform event visual cue extraction to obtain a compact description of the main visual event:

$$s_v = f_{\text{VLM}}(I, P_{\text{evt}}), \quad (2)$$

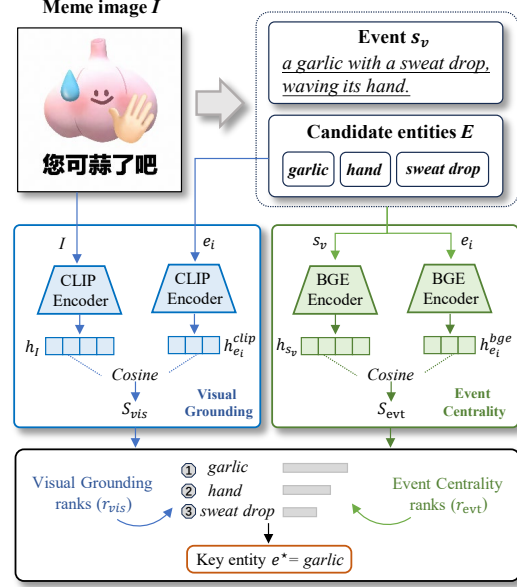


Figure 5: Overview of key visual entity selection.

where s_v describes the main action, object state, or scene-level relation in the meme image. While entity-level cues answer what visual elements are present, the event-level cue captures how these elements appear or interact in the visual context.

To ensure that the extracted cues reflect image-side evidence, the visual cue extraction process focuses on visible content and avoids transcribing overlaid captions, subtitles, watermarks, or other textual elements.

4.2 Key Visual Entity Selection

After visual semantic cue extraction, the candidate entity set E may contain peripheral entities, generic descriptions, or entities that are weakly related to pun interpretation. Therefore, as shown in Fig. 5, we further select a key visual entity from E . Instead of relying on the generation order of the vision-language model, we rank candidate entities from two complementary perspectives: visual grounding and event centrality.

Visual Grounding. Visual grounding measures how strongly each candidate entity is supported by the meme image. For each entity $e_i \in E$, we use CLIP (Yang et al., 2022) to compute the similarity between the image I and the textual representation of e_i . Specifically, the image and the entity text are encoded into a shared embedding space:

$$\mathbf{h}_I^{\text{clip}} = f_I^{\text{clip}}(I), \mathbf{h}_{e_i}^{\text{clip}} = f_T^{\text{clip}}(e_i), \quad (3)$$

where $f_I(\cdot)$ and $f_T(\cdot)$ denote the image encoder and text encoder of CLIP, respectively. The visual

Model	Method	Pun Meme Detection			Pun Meme Sentiment Analysis		
		TPR	TNR	F1	Acc _{target}	Acc _{type}	Acc _{overall}
<i>Closed-source</i>							
GPT-5.5	w/o Visual Cues	0.9928	0.6835	0.7562	0.6450	0.7592	0.4983
	+ Visual Cues	0.9856	0.7122	0.7697	0.6608	0.7742	0.5217
Gemini-3.1-Pro-Preview	w/o Visual Cues	0.9784	0.6079	0.7083	0.6692	0.8125	0.5542
	+ Visual Cues	0.9712	0.6511	0.7278	0.6867	0.8233	0.5758
<i>Open-source</i>							
LLaVA-1.5	w/o Visual Cues	0.7194	0.9568	0.7968	0.5108	0.4675	0.2708
	Ours	0.7842	0.9676	0.8482	0.5033	0.5517	0.3125
MiniCPM-V 2.6	w/o Visual Cues	0.8345	0.9245	0.8406	0.5867	0.5917	0.3308
	Ours	0.8561	0.9317	0.8592	0.6233	0.5867	0.3733
Qwen3-VL	w/o Visual Cues	0.8273	0.9460	0.8550	0.4892	0.6825	0.3767
	Ours	0.8561	0.9496	0.8750	0.5408	0.7142	0.4208
InternVL3	w/o Visual Cues	0.8058	0.9209	0.8205	0.5575	0.6450	0.3708
	Ours	0.8417	0.9101	0.8327	0.5692	0.6708	0.4100
GLM-4.1V	w/o Visual Cues	0.8777	0.9317	0.8714	0.5383	0.5367	0.3408
	Ours	0.8777	0.9460	0.8841	0.5400	0.5800	0.3567

Table 1: Main results on pun meme detection and sentiment analysis.

grounding score is defined as:

$$S_{\text{vis}}(e_i) = \cos(\mathbf{h}_I^{\text{clip}}, \mathbf{h}_{e_i}^{\text{clip}}), \quad (4)$$

A higher $S_{\text{vis}}(e_i)$ means that the candidate entity is more strongly grounded in the visual content of the meme image. This score helps distinguish visually supported entities from entities that are peripheral or weakly grounded in the image.

Event Centrality. Visual grounding alone cannot determine whether a candidate entity is central to the visual semantics of the meme. Some entities may be clearly visible but only serve as background or peripheral elements. Therefore, we further measure the semantic association between each candidate entity and the event-level cue s_v .

Given s_v and a candidate entity $e_i \in E$, we use BGE (Xiao et al., 2023) to encode them into the same semantic embedding space:

$$\mathbf{h}_{s_v}^{\text{bge}} = g^{\text{bge}}(s_v), \mathbf{h}_{e_i}^{\text{bge}} = g^{\text{bge}}(e_i), \quad (5)$$

where $g(\cdot)$ denotes the BGE encoder. The event centrality score is computed as:

$$S_{\text{evt}}(e_i) = \cos(\mathbf{h}_{s_v}^{\text{bge}}, \mathbf{h}_{e_i}^{\text{bge}}), \quad (6)$$

A higher $S_{\text{evt}}(e_i)$ indicates that the entity is more closely related to the action, state, or scene relation described by the event-level cue. This score complements visual grounding by identifying entities that are not only visible but also semantically central to the image-derived event.

Rank-based Entity Selection. Visual grounding and event centrality characterize candidate entities from image-level visibility and event-level relevance, respectively. To obtain an entity that is consistently supported by both perspectives, we employ a rank-based aggregation strategy inspired by Borda Count.

For each scoring perspective $c \in \{\text{vis}, \text{evt}\}$, all candidate entities are ranked in descending order according to their scores. Let $r_c(e_i) \in \{1, \dots, m\}$ denote the rank of entity e_i under c , where a smaller value indicates a higher rank. The rank-based score of each candidate entity is computed as:

$$B(e_i) = \sum_{c \in \{\text{vis}, \text{evt}\}} (m - r_c(e_i)), \quad (7)$$

The final key entity is selected as:

$$e^* = \arg \max_{e_i \in E} B(e_i), \quad (8)$$

In this way, the selected key entity is expected to be both visually grounded in the image and semantically central to the extracted event-level cue.

4.3 Cue-conditioned Joint Prediction

Given the meme image I , caption C , selected key entity e^* , and event-level cue s_v , we construct a cue-conditioned input X and generate the structured joint prediction Y :

$$\begin{aligned} X &= [I; C; e^*; s_v; P_{\text{joint}}], \\ Y &= (y_{\text{pun}}, y_{\text{target}}, y_{\text{type}}). \end{aligned} \quad (9)$$

Hyperparameter	Value
Seed	42
LoRA rank	8
Epochs	3
Batch size	1
Gradient accumulation	8
Learning rate	1e-4
LR scheduler	Cosine
Optimizer	AdamW

Table 2: Main hyperparameter settings.

where P_{joint} specifies the joint prediction instruction and structured output format. The sentiment fields are set to null for non-pun memes.

We perform LoRA-based supervised fine-tuning by minimizing the token-level negative log-likelihood of the structured output:

$$\mathcal{L} = - \sum_{t=1}^{|Y|} \log P(y_t \mid y_{<t}, X), \quad (10)$$

Thus, pun detection and sentiment analysis are jointly modeled under the same visually cue-conditioned context. Prompt templates are provided in Appendix A.

5 Experiments

Implementation Details We implement the open-source experiments using LoRA-based supervised fine-tuning. For each open-source VLM, we follow the same training protocol and use the hyperparameter settings reported in Table 2. We use Qwen3-VL-8B-Instruct as a separate frozen visual cue extractor for all cue-enhanced settings. The extractor is not fine-tuned on PUNMEMECN and is shared across all evaluated models. The open-source VLM backbones are fine-tuned separately for final prediction. All open-source fine-tuning experiments are conducted on a single NVIDIA RTX 3090 GPU. During training, we use BF16 precision to reduce memory consumption and keep the batch size fixed across different backbones for fair comparison.

Dataset We conduct experiments on PUNMEMECN (Xu et al., 2025), which contains 1,959 meme images, including 653 pun memes and 1,306 non-pun memes. Each pun meme is paired with two non-pun distractors, i.e., one image-similar sample and one caption-similar sample. The dataset provides captions and annotations for pun

meme detection and sentiment analysis, where sentiment labels consist of target and type. We use an approximately 70/10/20 group-level split for training, validation, and testing, ensuring that each pun meme and its distractors are assigned to the same subset to avoid information leakage. For sentiment evaluation, following the sampling strategy adopted in PUNMEMECN, we apply target-balanced upsampling only to the gold pun memes in the test set, sampling each sentiment-target category to 400 instances with a fixed random seed. We evaluate the models on two subtasks. For pun meme detection, we report TPR, TNR, and F1-score (F1). For pun meme sentiment analysis, we report $\text{Acc}_{\text{target}}$, Acc_{type} , and $\text{Acc}_{\text{overall}}$, which measure target prediction, sentiment type prediction, and exact matching of both dimensions, respectively.

Baselines We compare our framework with both closed-source and open-source VLMs. For closed-source models, we evaluate **GPT-5.5** (OpenAI, 2026) and **Gemini-3.1-Pro-Preview** (Google DeepMind, 2026) under zero-shot settings, comparing their performance with and without the proposed visual cues. For open-source models, we evaluate several representative VLM backbones, including **LLaVA-1.5-7B-HF** (Liu et al., 2024), **MiniCPM-V 2.6** (Yao et al., 2024), **Qwen3-VL-8B-Instruct** (Qwen, 2025), **InternVL3-8B** (Zhu et al., 2025), and **GLM-4.1V-9B-Thinking**¹. For each open-source backbone, we compare two variants under the same fine-tuning setting: *w/o Visual Cues*, which uses only the meme image and caption, and *Ours*, which additionally incorporates the selected key visual entity and event-level cue. These comparisons evaluate whether the proposed structured visual cues provide consistent improvements across different VLM backbones and model settings.

5.1 Main Results

Table 1 reports the main results on pun meme detection and sentiment analysis across both open-source and closed-source VLMs. For the open-source backbones, *Ours* consistently improves F1 over the corresponding *w/o Visual Cues* setting. Similar improvements are observed for the closed-source VLMs under zero-shot evaluation, indicating that the proposed visual semantic cues provide

¹<https://huggingface.co/zai-org/GLM-4.1V-9B-Thinking>

Method	Pun	Sentiment
Ours	0.8750	0.4208
- Event-level Cue	0.8571	0.3825
- Key Entity	0.8657	0.3950
w/o Visual Cues	0.8550	0.3767

Table 3: Ablation results. The **Pun** and **Sentiment** columns report F1 for pun meme detection and overall accuracy for pun meme sentiment analysis, respectively.

complementary information across different model settings.

For the closed-source VLMs, both GPT-5.5 and Gemini-3.1-Pro-Preview exhibit high TPR but relatively lower TNR, consistent with the pun-prediction bias observed in PUNMEMECN. Incorporating the structured visual cues mainly improves non-pun recognition while preserving strong pun detection performance, leading to higher F1 scores. For the open-source backbones, the proposed method also yields consistent F1 improvements, demonstrating that the effectiveness of the visual cues is not restricted to a particular VLM architecture.

The improvements are particularly evident in sentiment analysis. $\text{Acc}_{\text{overall}}$ increases across all evaluated models, with absolute gains ranging from 1.59 to 4.41 percentage points. This indicates that the selected key visual entity and event-level cue provide useful contextual information for jointly identifying the sentiment target and sentiment type. Although a few individual metrics show slight decreases, F1 and $\text{Acc}_{\text{overall}}$ consistently improve across all evaluated models. Overall, these results demonstrate that structured visual semantic cues provide reliable complementary information for pun meme understanding, with statistically significant improvements in the main overall metrics, as assessed by paired bootstrap tests ($p < 0.05$).

5.2 Ablation Study

To examine the contributions of the two visual semantic cues, we conduct ablation experiments using Qwen3-VL as the representative backbone. Starting from the full model, we remove either the event-level cue (- *Event-level Cue*) or the selected key entity (- *Key Entity*) from the cue-conditioned prediction input, or remove both cues (*w/o Visual Cues*). As shown in Table 3, the full model achieves the best performance on both pun detection and sentiment analysis. Removing either cue leads to performance degradation, demonstrating that the

Method	Pun	Sentiment
Single Entity + Joint Extraction	0.8399	0.3735
Single Entity + Separate Extraction	0.8460	0.3829
Multiple Entities + Joint Extraction	0.8540	0.4017
Ours	0.8750	0.4208

Table 4: Comparison of visual semantic cue extraction strategies.

Method	Pun	Sentiment
Ours	0.8750	0.4208
- Visual Grounding	0.8551	0.4000
- Event Centrality	0.8489	0.4067
Random Selection	0.8470	0.3808

Table 5: Effect of different key entity selection strategies.

key entity and event-level cue provide complementary information for pun meme understanding. In particular, removing the event-level cue results in a larger performance drop, suggesting that contextual event semantics play an important role in capturing the visual context associated with pun interpretation, while the selected key entity provides complementary grounding for caption-induced ambiguity. Unless otherwise noted, Qwen3-VL is used as the default backbone in the following experiments.

5.3 Comparison of Cue Extraction Strategies

To examine the design choices in visual semantic cue extraction, we compare different entity extraction and cue extraction strategies. As shown in Table 4, *Single Entity* denotes directly extracting one visual entity, whereas *Multiple Entities* denotes extracting multiple entities followed by key entity selection. Meanwhile, *Joint Extraction* and *Separate Extraction* indicate whether entity-level and event-level cues are extracted in a single process or through two dedicated processes, respectively.

Our *Multiple Entities + Separate Extraction* strategy achieves the best performance, with an F1 of 0.8750 and an $\text{Acc}_{\text{overall}}$ of 0.4208. The results show that retaining multiple entities before selection avoids premature entity decisions, while separate extraction better preserves the complementary roles of entity-level and event-level cues. These results support extracting multiple entities separately from the event-level cue before key entity selection.

5.4 Effect of Key Entity Selection

To analyze the effect of key entity selection, we compare different strategies for selecting one entity

Case Type	Input	Meanings	Prediction Comparison			
			Setting	Meme type	Sentiment target	Sentiment type
(A) Joint Correction	 我梳了 我梳了	Literal : <i>I combed</i> Intended: <i>I lost</i>  ↔ 梳 (comb) ↔ 输 (lose) Visual Cues : <i>comb; Rabbit and comb combination</i>	Gold	<i>pun</i>	<i>self</i>	<i>neutral</i>
			w/o Cues	<i>non-pun</i>	<i>null</i>	<i>null</i> ✗
			Ours	<i>pun</i>	<i>self</i>	<i>neutral</i> ✓
(B) Joint Correction	 这像画吗 这像画吗	Literal : <i>Does this look like a painting?</i> Intended: <i>Does this make any sense?</i>  ↔ 画 (painting) ↔ 话 (sense) Visual Cues : <i>picture frame; A dog's head is framed in a picture frame.</i>	Gold	<i>pun</i>	<i>other</i>	<i>neutral</i>
			w/o Cues	<i>non-pun</i>	<i>null</i>	<i>null</i> ✗
			Ours	<i>pun</i>	<i>other</i>	<i>neutral</i> ✓
(C) Sentiment Refinement	 目瞪口呆 目瞪口呆	Literal : <i>staring wide-eyed at a pocket</i> Intended: <i>be dumbfounded</i>  ↔ 袋 (pocket) ↔ 呆 (dumbfounded) Visual Cues : <i>plastic bag; A hand covering a plastic bag with a water droplet on it.</i>	Gold	<i>pun</i>	<i>self</i>	<i>negative</i>
			w/o Cues	<i>pun</i>	<i>other</i>	<i>neutral</i> ✗
			Ours	<i>pun</i>	<i>self</i>	<i>negative</i> ✓

Figure 6: Examples of case study.

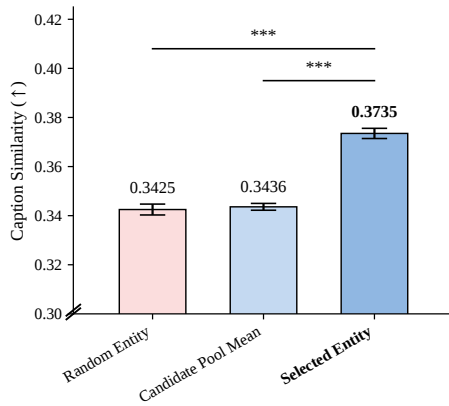


Figure 7: Entity-caption association analysis.

from the candidate entity set E . As shown in Table 5, our rank-based aggregation strategy achieves the best performance on both F1 and $\text{Acc}_{\text{overall}}$. Compared with random selection, our strategy substantially improves the results, indicating that selecting an arbitrary candidate entity cannot provide stable visual evidence for pun interpretation. Removing either visual grounding or event centrality leads to performance degradation. This suggests that visual grounding and event centrality capture different aspects of entity relevance: the former measures whether an entity is visually salient in the image, while the latter evaluates whether it is semantically central to the event-level cue. These results show that combining both perspectives is beneficial for selecting pun-relevant visual entities.

5.5 Entity-Caption Association Analysis

We further examine whether the selected entities establish meaningful semantic associations with the accompanying captions. Specifically, we use BGE (Xiao et al., 2023) to compute the similarity between each caption and the selected entity, and compare it with a randomly sampled entity from the same candidate set and the average similarity over all candidates in that set.

As shown in Fig. 7, the selected entities achieve a higher average caption similarity (0.3735) than both randomly sampled entities (0.3425) and the candidate-pool mean (0.3436). The improvements over both baselines are statistically significant under paired tests ($p < 0.001$). These results suggest that the selected visual entities provide more relevant image-side evidence for connecting visual semantics with the accompanying captions, thereby supporting subsequent multimodal reasoning for pun meme understanding.

5.6 Case Study

To qualitatively analyze the effect of visual semantic cues, Fig. 6 presents representative cases with their literal and intended meanings, together with predictions under the joint output format of pun detection, sentiment target, and sentiment type. Cases (A) and (B) show *joint correction*. In Case (A), the extracted entity highlights “梳” in “我梳了”, helping the model associate it with the homophonic reading “我输了” (*I lost*). In Case (B), the framed picture makes “画” salient in “这像画吗”, which,

together with the familiar expression “这像话吗” (*Does this make any sense?*), supports correct pun and sentiment predictions. Case (C) illustrates *sentiment refinement*. Although both variants identify “目瞪口呆” as a pun, the pocket, sweat drop, and hand-covering-mouth gesture provide richer affective evidence, enabling our model to correctly interpret its relation to “目瞪口呆” (*be dumbfounded*) and refine the sentiment prediction.

Overall, these cases demonstrate that visual semantic cues facilitate the interpretation of visually grounded ambiguity and improve both pun detection and sentiment understanding across representative examples.

6 Conclusion

In this paper, we address pun meme understanding as a challenging multimodal reasoning problem, where implicit humor and sentiment emerge from the interaction between caption-induced ambiguity and visual semantics. To move beyond holistic image-caption reasoning, we introduce multi-source visual semantic cues as explicit complementary evidence and incorporate them into the joint modeling of pun meme detection and sentiment analysis. This cue-enhanced formulation enables more precise modeling of the fine-grained cross-modal correlations underlying pun interpretation. Comprehensive experiments demonstrate that the proposed method consistently improves overall performance across multiple backbone architectures. Overall, this work highlights the importance of explicitly modeling fine-grained visual semantic evidence for understanding implicit humor and affect in multimodal memes, and offers a meaningful perspective on reasoning over semantically ambiguous multimodal content.

Limitations

Pun memes are often shaped by specific cultural contexts, where the interpretation of a caption and its connection to visual content may rely on shared background knowledge, familiar expressions, and culturally recognizable forms of humor. The present study focuses on the meme context covered by the evaluated benchmark and investigates how explicit visual semantic cues can support pun meme understanding within this setting. In broader cross-cultural scenarios, ambiguous captions may interact with visual evidence in different ways, and similar visual elements may convey dif-

ferent humorous or affective implications across communities. Future research may further explore pun meme understanding in more diverse cultural contexts.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62376178), Jiangsu Key Laboratory of Language Computing (JSLCKeyLab 202500003), and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

- Christian Bauckhage. 2011. [Insights into internet memes](#). In *Proceedings of the Fifth International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July 17-21, 2011*. The AAAI Press.
- Jiwan Chung, Seungwon Lim, Jaehyun Jeon, Seungbeen Lee, and Youngjae Yu. 2024. [Can visual language models resolve textual ambiguity with visual cues? let visual puns tell you!](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 2452–2469. Association for Computational Linguistics.
- Anastasia Denisova. 2019. *Internet memes and society: Social, cultural, and political contexts*. Routledge.
- Google DeepMind. 2026. Gemini 3.1 pro model card. <https://deepmind.google/models/model-cards/gemini-3-1-pro/>. Accessed 2026.
- Eemeli Hakoköngäs, Otto Halmesvaara, and Inari Sakki. 2020. Persuasion through bitter humor: Multimodal discourse analysis of rhetoric in internet memes of two far-right groups in finland. *Social media+ society*, 6(2):2056305120921575.
- Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. [Decoding the underlying meaning of multimodal hateful memes](#). In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China*, pages 5995–6003. ijcai.org.
- Eftekhari Hossain, Omar Sharif, and Mohammed Moshul Hoque. 2022. [Memosen: A multimodal dataset for sentiment analysis of memes](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 1542–1554. European Language Resources Association.
- Justine T. Kao, Roger Levy, and Noah D. Goodman. 2016. [A computational model of linguistic humor in puns](#). *Cogn. Sci.*, 40(5):1270–1285.

- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. [The hateful memes challenge: Detecting hate speech in multimodal memes](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Liisi Laineste and Piret Voolaid. 2016. Laughing across borders: Intertextuality of internet memes. *The European Journal of Humour Research*, 4(4):26–49.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. [Improved baselines with visual instruction tuning](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.
- Yiran Rex Ma, Shan Huang, Yuting Xu, Ziyu Zhou, and Yuanxi Wei. 2025. [Pun2pun: Benchmarking llms on textual-visual chinese-english pun translation via pragmatics model and linguistic reasoning](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 331–354. Association for Computational Linguistics.
- Tristan Miller and Iryna Gurevych. 2015. [Automatic disambiguation of english puns](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 719–729. The Association for Computer Linguistics.
- Tristan Miller and Mladen Turković. 2016. Towards the automatic detection and identification of english puns. *The European Journal of Humour Research*, 4(1):59–75.
- OpenAI. 2026. Gpt-5.5 system card. <https://openai.com/index/gpt-5-5-system-card/>. Accessed 2026.
- Qwen. 2025. [Qwen3-vl technical report](#). *CoRR*, abs/2511.21631.
- Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. [Semeval-2020 task 8: Memotion analysis- the visuo-lingual metaphor!](#) In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 759–773. International Committee for Computational Linguistics.
- Maja Šimon. 2019. *Puns in Internet Memes: A Study of 9GAG*. Ph.D. thesis.
- Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022. [Expunations: Augmenting puns with keywords and explanations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 4590–4605. Association for Computational Linguistics.
- Shardul Suryawanshi, Bharathi Raja Chakravarthi, Michael Arcan, and Paul Buitelaar. 2020. [Multimodal meme dataset \(multioff\) for identifying offensive content in image and text](#). In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, TRAC@LREC 2020, Marseille, France, May 2020*, pages 32–41. European Language Resources Association (ELRA).
- Yufei Tian, Divyanshu Sheth, and Nanyun Peng. 2022. [A unified framework for pun generation with humor principles](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Findings of ACL, pages 3253–3261. Association for Computational Linguistics.
- Camilla Vásquez and Erhan Aslan. 2021. “cats be outside, how about meow”: multimodal humor and creativity in an internet meme. *Journal of Pragmatics*, 171:101–117.
- Bingbing Wang, Shijue Huang, Bin Liang, Geng Tu, Min Yang, and Ruifeng Xu. 2024. [What do they "meme"? A metaphor-aware multi-modal multi-task framework for fine-grained meme understanding](#). *Knowl. Based Syst.*, 294:111778.
- Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. [C-pack: Packaged resources to advance general chinese embedding](#). *CoRR*, abs/2309.07597.
- Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. ["a good pun is its own reword": Can large language models understand puns?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 11766–11782. Association for Computational Linguistics.
- Zhijun Xu, Siyu Yuan, Yiqiao Zhang, Jingyu Sun, Tong Zheng, and Deqing Yang. 2025. [Punmemecn: A benchmark to explore vision-language models’ understanding of chinese pun memes](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 18694–18710. Association for Computational Linguistics.
- An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. [Chinese CLIP: contrastive vision-language pretraining in chinese](#). *CoRR*, abs/2211.01335.

- Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, and 4 others. 2024. [Minicpm-v: A GPT-4V level MLLM on your phone](#). *CoRR*, abs/2408.01800.
- Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. [A neural approach to pun generation](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1650–1660. Association for Computational Linguistics.
- Alessandro Zangari, Matteo Marcuzzo, Andrea Albarelli, Mohammad Taher Pilehvar, and José Camacho-Collados. 2025. [Pun unintended: Llms and the illusion of humor understanding](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 27936–27971. Association for Computational Linguistics.
- Tuo Zhang, Tiantian Feng, Yibin Ni, Mengqin Cao, Ruying Liu, Kiana Avestimehr, Katharine Butler, Yanjun Weng, Mi Zhang, Shrikanth Narayanan, and Salman Avestimehr. 2025. [Creating a lens of chinese culture: A multimodal dataset for chinese pun rebus art understanding](#). In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, Findings of ACL, pages 22473–22487. Association for Computational Linguistics.
- Li Zheng, Hao Fei, Ting Dai, Zuquan Peng, Fei Li, Huisheng Ma, Chong Teng, and Donghong Ji. 2025. [Multi-granular multimodal clue fusion for meme understanding](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 26057–26065. AAAI Press.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. [Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models](#). *CoRR*, abs/2504.10479.
- Yanyan Zou and Wei Lu. 2019. [Joint detection and location of english puns](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 2117–2123. Association for Computational Linguistics.

A Prompt Templates

Figures 8 and 9 present the prompt templates for visual semantic cue extraction and cue-conditioned joint prediction, respectively.

Prompts for Visual Semantic Cue Extraction

Input: <meme image>

P_{ent} : Entity-level Cue Extraction
Instruction: Please analyze the meme image and extract the visible entities. Do not transcribe subtitles or watermarks in the image. Only focus on concrete visual content.

Output Format:
Entities: XXX; YYY; ZZZ

P_{evt} : Event-level Cue Extraction
Instruction: Please analyze the meme image and describe the main visual event, state, or action. Do not transcribe captions, subtitles, texts, or watermarks in the image.

Output Format:
Event: XXX

Figure 8: Prompt templates for visual semantic cue extraction.

Prompt for Joint Learning

Input of Meme Image: <meme image>

Definition:
A meme is a “pun meme” if its caption intentionally uses polysemy or homophony to create two or more meanings, and at least one meaning is grounded in the visual elements of the image. Otherwise, it is a “non-pun meme”.

Task Description:
Based on the above definition, determine whether the given meme is a pun meme. The caption within the meme is “{caption}”. The visual elements are “Key Entity: {key_entity}; Event: {event}”. Pay attention to whether there is a homophonic or polysemous pun association between the caption and the image. If a pun exists, both the literal meaning and another implicit meaning should be satisfied. If it is a pun meme, further predict the sentiment it expresses in online chatting, including the sentiment target and sentiment type. If it is not a pun meme, set sentiment analysis to null.

Output Format:
{
 “pun_label”: “pun/non-pun”,
 “sentiment_analysis”: {
 “sentiment_target”: “self/both/other”,
 “sentiment_type”: “positive/neutral/negative”
 } or “null”
}

Your response:

Figure 9: The cue-conditioned prompt template used for joint pun meme detection and sentiment analysis.