

SatireDecoder: Visual Cascaded Decoupling for Enhancing Satirical Image Comprehension

Yue Jiang^{1*}, Haiwei Xue^{2, 5, 6*}, Minghao Han¹, Mingcheng Li¹, Xiaolu Hou¹,
Dingkang Yang^{1,3†}, Lihua Zhang^{1†}, Xu Zheng^{4, 5, 6†}

¹College of Intelligent Robotics and Advanced Manufacturing, Fudan University

²Tsinghua University

³ByteDance

⁴INSAIT, Sofia University “St. Kliment Ohridski”

⁵The Hong Kong University of Science and Technology

⁶The Hong Kong University of Science and Technology (Guangzhou)

jiangyue23@m.fudan.edu.cn, winniygd@outlook.com,

{lihuazhang, dkyang20}@fudan.edu.cn, zhengxu128@gmail.com

Abstract

Satire, a form of artistic expression combining humor with implicit critique, holds significant social value by illuminating societal issues. Despite its cultural and societal significance, satire comprehension, particularly in purely visual forms, remains a challenging task for current vision-language models. This task requires not only detecting satire but also deciphering its nuanced meaning and identifying the implicated entities. Existing models often fail to effectively integrate local entity relationships with global context, leading to misinterpretation, comprehension biases, and hallucinations. To address these limitations, we propose SatireDecoder, a training-free framework designed to enhance satirical image comprehension. Our approach proposes a multi-agent system performing visual cascaded decoupling to decompose images into fine-grained local and global semantic representations. In addition, we introduce a chain-of-thought reasoning strategy guided by uncertainty analysis, which breaks down the complex satire comprehension process into sequential sub-tasks with minimized uncertainty. Our method significantly improves interpretive accuracy while reducing hallucinations. Experimental results validate that SatireDecoder outperforms existing baselines in comprehending visual satire, offering a promising direction for vision-language reasoning in nuanced, high-level semantic tasks.

Introduction

Satirical images often rely on deliberately contradictory or conflicting scenes to convey the deep semantics, blending humor with subtle critique (del Pilar Salas-Zárate et al. 2017). People frequently employ satirical images on social media platforms to express their attitudes toward social phenomena or trending events. Consequently, comprehending the satirical semantics inherent in images holds considerable scholarly significance. Understanding satirical images

requires identifying inherent conflicts and analyzing the interaction between local entities and global contexts to infer the deep semantics.

As illustrated in Figure 1, previous works (Yu et al. 2024; Zhu et al. 2024b; Xie et al. 2024; Yue et al. 2024) have exclusively focused on simplistic satire detection, which is an easy binary classification, neglecting the more challenging task of understanding and interpreting the deep satirical semantics inherent in images. Due to the difference between the image’s satirical and surface meanings, comprehending its deep semantics necessitates a thorough analysis of the relationships between local entities and the global context to uncover contradictory or incongruent components. Moreover, previous methods (Wu 2018; Chen et al. 2024a,c) has depended on large datasets and high training costs, thereby suffering from substantial overhead and limited portability. In addition, despite the remarkable success of current multimodal large language models (MLLMs) in multimodal tasks, several popular MLLMs exhibit significant limitations in comprehending image deep semantics beyond the surface meanings (Chang et al. 2024). MLLMs tend to overlook or fabricate local entities and crucial details within images (Chen et al. 2024b,d; Yang et al. 2024b,a), resulting in hallucination issues and the misinterpretation of satirical semantics (Leng et al. 2024). Furthermore, MLLMs lack a step-by-step inference process from local entities to global context during image understanding, creating significant challenges in grasping the relationships between visual elements and the deep semantics of satire (Huang et al. 2025).

To address the limitations of MLLMs in comprehending satirical images, we propose SatireDecoder, a novel training-free framework illustrated in Figure 1. SatireDecoder employs a multi-agent visual cascaded decoupling mechanism to decompose images into fine-grained semantic representations, effectively capturing both local entity features and global contextual cues. This design further enables the identification of semantic discrepancies, contradic-

*These authors contributed equally.

†Corresponding authors.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

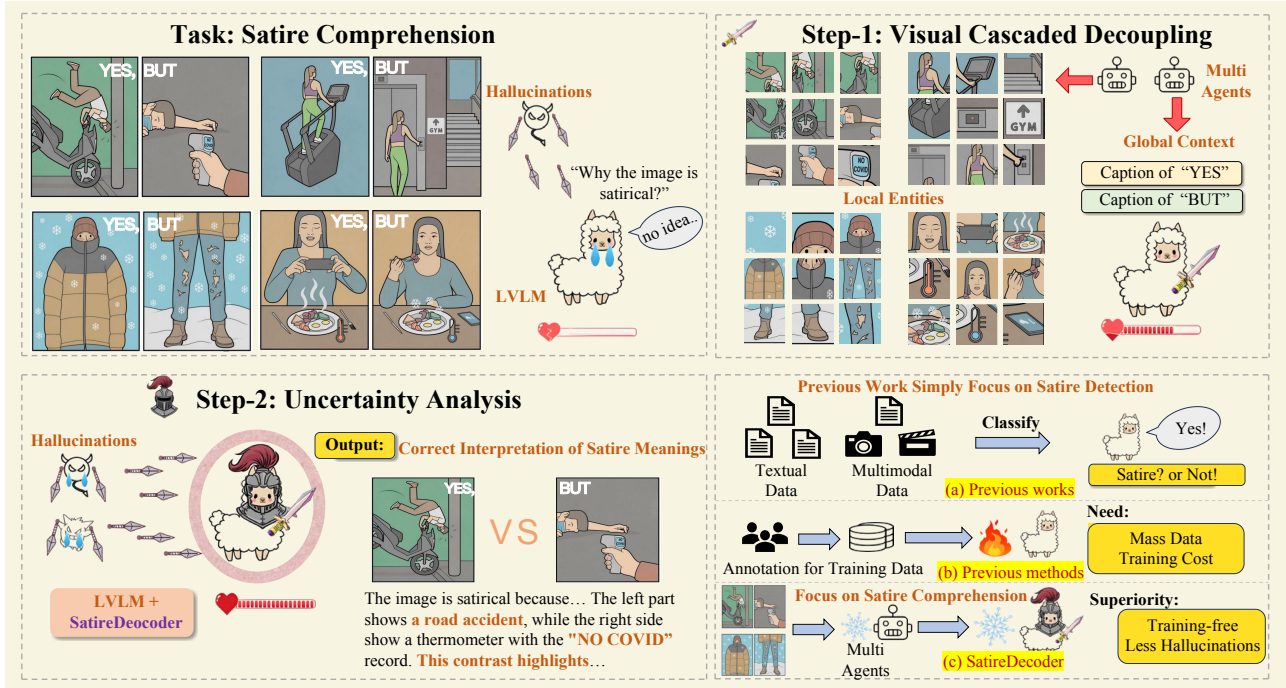


Figure 1: SatireDecoder consists of a multi-agent visual cascaded decoupling module and an uncertainty analysis strategy, which together help MLLMs capture discrepancies between local entities and the global context, thereby enhancing models’ ability to comprehend satirical images.

tions, and incongruities—key characteristics of visual satire. Subsequently, a Chain-of-Thought (CoT) reasoning strategy guided by uncertainty analysis decomposes the complex satire comprehension process into sequential subtasks with minimized uncertainty, improving interpretability while mitigating hallucinations. SatireDecoder can be seamlessly integrated into various MLLM baselines and consistently enhances their performance in satire understanding. Extensive experiments and ablation studies further validate the effectiveness of SatireDecoder and the contribution of each component within the framework. Our main contributions are summarized as follows:

- We propose SatireDecoder, a novel training-free framework that leverages multi-agent collaboration for visual cascaded decoupling, decomposing complex satirical images into fine-grained semantic representations across local and global levels, enhancing visual perception.
- We design a CoT reasoning, guiding MLLMs to decompose the satire comprehension into three subtasks: identifying local entities, understanding global context, and inferring satirical intent. This approach improves interpretability and demonstrates generalization ability in complex, resource-constrained visual reasoning tasks.
- We introduce an uncertainty-guided inference method, which quantifies the discrepancy between the outputs of MLLMs and multi-agents for shared subtasks. By minimizing the uncertainty score, our method reduces hallu-

inations and enhances the robustness of final satire interpretation, offering a novel perspective on interpretable model optimization through uncertainty analysis.

Related Work

Text-based Satire Analysis

Satire comprehension is crucial for sentiment analysis and identifying harmful comments. Previous studies on satire in text modality focus on the satire detection (Joshi, Bhattacharyya, and Carman 2017). Several supervised approaches are applied to the tasks, including traditional machine learning methods with lexical features (Ptáček, Habernal, and Hong 2014; Bouazizi and Ohtsuki 2015) and deep learning method (Wu 2018). Moreover, various emotional (Thu and Aung 2018; Li et al. 2024), psychological (del Pilar Salas-Zárate et al. 2017), and linguistic features (Yang, Mukherjee, and Dragut 2017) are incorporated to enhance satire detection. The effect of contextual inconsistencies in satire detection has also been explored (Joshi, Sharma, and Bhattacharyya 2015), emphasizing the importance of semantic and pragmatic factors. However, images have emerged as a pivotal medium for information dissemination on social media. Text-based satire analysis is inadequate for fully capturing the satirical content present in contemporary media.

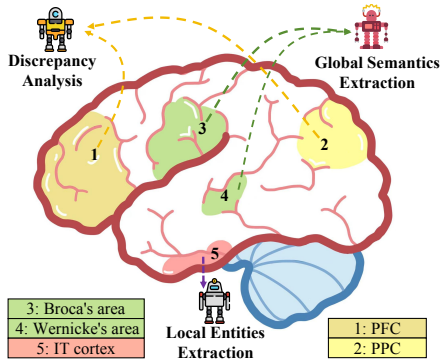


Figure 2: In the multi-agent system, the Local Entities Extraction Agent simulates the InferoTemporal cortex (IT cortex) (Grill-Spector and Malach 2004), while the Global Semantics Extraction Agent simulates the Broca’s area and the Wernicke’s area (Jäncke, Liem, and Merillat 2021). Additionally, the Discrepancy Analysis Agent imitates the function of the PreFrontal Cortex (PFC) and the Posterior Parietal Cortex (PPC) (Grill-Spector and Malach 2004).

Multimodal Satire Analysis

Multimodal satire presents unique challenges due to the interplay between modalities. Previous research has focused primarily on detection and classification tasks (Cai 2019; Castro et al. 2019; Yu et al. 2024; Zhu et al. 2024b; Xie et al. 2024; Xue et al. 2025). Early approaches (Schifanella et al. 2016; Das and Clark 2018) employ traditional computer vision techniques combined with text analysis for meme classification. Recent methods have evolved toward deep learning architectures, incorporating pre-trained vision and language models for enhanced feature extraction (Cai 2019; Bharti and Gupta 2022). However, these approaches typically treat satire detection as a binary classification task, without addressing the deeper understanding of satirical elements and their interactions. Satire comprehension involves identifying contextual inconsistencies and reasoning about the satirical intent, evaluated by how well the visual differences and motivations are captured.

In the general domain, numerous low-cost, training-free methods exist to enhance the multimodal comprehension capabilities of MLLMs. Several works utilize multi-agent systems (Li et al. 2025; Jiang et al. 2024) or incorporate specialized external models (Zeng et al. 2023; Hyun et al. 2024) to augment multimodal processing. Some approaches (Jiang et al. 2025a; Bi et al. 2025b,a; Zhao et al. 2025) further integrate chain-of-thought reasoning to guide the model in decomposing understanding into sequential steps. Regarding the mitigation of hallucinations in MLLMs, mainstream methods (Leng et al. 2024; Zhang et al. 2025; Wang et al. 2025, 2024) involve optimizing the model during decoding by applying principles of contrastive learning, among others.

Datasets for Satire Analysis

Beyond conventional text-based satire detection, several new datasets have been developed for multimodal scenes.

MMSD (Cai 2019), MMSD2.0 (Qin et al. 2023), MUSTARD (Castro et al. 2019), and SarcNet (Yue et al. 2024) focus solely on satire detection, failing to assess the ability of MLLMs to comprehend the deep semantics of satire. NYK-MS (Chang et al. 2024) consists of more than 1,500 cartoon-caption pairs, supporting both satire detection and comprehension. But in NYK-MS, satire is conveyed through the combination of image and textual descriptions, which simplifies the comprehension task. In our paper, we utilize YesBut (Nandy et al. 2024), which is the only current dataset specifically designed for satire comprehension and without image captions. Each satirical image in YesBut is structured in a “Yes, But” format, where the left half of the image depicts a normal scene, juxtaposed with a conflicting scene on the right, which together constitute the deep semantics of satire. In YesBut, satire is conveyed solely through visual information, without accompanying any textual clues, posing a unique challenge for MLLMs.

Methodology

The framework of SatireDecoder is depicted in Figure 3. Inspired by the human perception paradigm on visual information (Yang et al. 2023; Mischler 2024) and the multi-agent collaboration (Jiang et al. 2025b; Li et al. 2025), we propose a multi-agent system for visual cascaded decoupling to decompose the visual input into fine-grained representations. Based on the multi-agent collaboration, a CoT prompt is constructed, which is combined with the fine-grained information from the image. Catalyzed by the CoT prompt, MLLM decomposed the complex satire comprehension task into three subtasks, followed by the uncertainty analysis strategy to minimize the uncertainty score during the three-step inference, which efficiently mitigates hallucinations.

Multi-agent-based Visual Cascaded Decoupling

Related studies (Mischler 2024; Bullier 2001; Grill-Spector and Malach 2004) have demonstrated that distinct regions of the cerebral cortex exhibit functional specificity in visual information processing. Inspired by this, we propose a biologically inspired multi-agent framework to simulate this mechanism. As shown in Figure 2, the agents are employed to play distinct roles, imitating different areas of the cerebral cortex. To extract local entity details, we employ the Local Entities Extraction Agent LE for image tagging, simulating IT cortex in the cerebral cortex, which includes object-selective regions and plays a crucial role in object recognition (Bullier 2001; Grill-Spector and Malach 2004). The satirical image consists of two parts, $\{Image_{yes}\}$ and $\{Image_{but}\}$, denoted as I_y and I_b respectively. The LE is utilized to effectively detect and assign relevant tags to local entities within I_y and I_b , and results are denoted as $LE_y = LE(I_y)$, and $LE_b = LE(I_b)$.

To grasp the global context, we leverage the Global Semantics Extraction Agent GS for image captioning, simulating PPC and PFC in the cerebral cortex. The two regions play essential roles in the integration of global visual information from complex scenes and the facilitation of high-level cognition and decision-making (Bullier 2001; Grill-Spector and Malach 2004). The GS is utilized to process

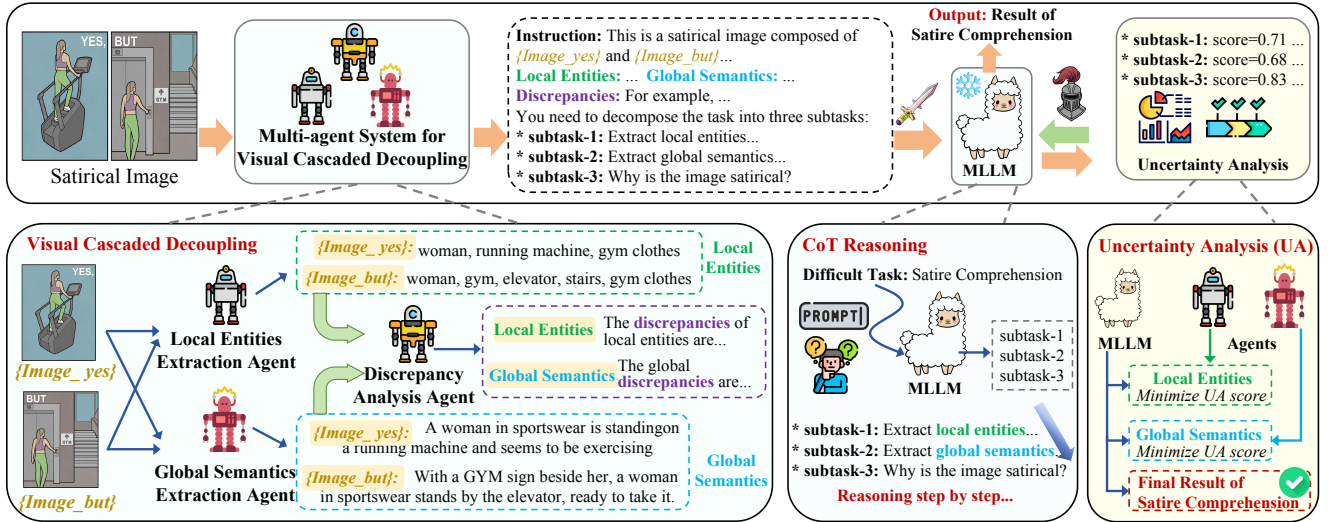


Figure 3: Framework of SatireDecoder. (1) Visual Cascaded Decoupling: A multi-agent system decouples the satirical image into fine-grained semantic representations. (2) Prompt Construction and CoT Reasoning: The fine-grained semantic representations form a structured prompt for CoT reasoning, breaking complex satire comprehension into three subtasks. (3) Inference Optimization with Uncertainty Analysis: progressively reduces uncertainty and hallucinations during reasoning.

the two contrasting scenes I_y and I_b , and the results can be represented as $GS_y = GS(I_y)$, and $GS_b = GS(I_b)$.

To contrast the subtle discrepancies and inconsistencies between the two scenes I_y and I_b depicted in the satirical image, we employed the Discrepancy Analysis Agent DA to simulate the Broca’s and Wernicke’s areas (Jäncke, Liem, and Merillat 2021) in the cerebral cortex, controlling complex vision-language comprehension (Bullier 2001; Grill-Spector and Malach 2004) and analyzing the outputs of LE and GS agents. The discrepancies of local entities D_l and global semantics D_g are represented as $D_l = DA(LE_y, LE_b)$ and $D_g = DA(GS_y, GS_b)$.

Within our multi-agent framework, to optimally balance cost and performance, we designate RAM (Zhang et al. 2024) and BLIP (Li et al. 2022) to play the role of Local Entities Extraction Agent and Global Semantics Extraction Agent, respectively. RAM is a specialized model for image tagging, engineered to accurately identify a wide range of common object categories within an input image, while BLIP is a pretraining model developed to bridge the gap between visual and linguistic understanding. It is capable of generating fluent and coherent descriptions that accurately reflect the content of an image. For the two elementary visual-semantic tasks, RAM and BLIP can match the performance of many MLLMs, paving the way for broader adoption of our training-free method. Concurrently, Qwen2 (QwenTeam et al. 2024) is employed as the Discrepancy Analysis Agent to undertake more complex, comprehension-intensive higher-level semantic tasks by harnessing the robust language understanding capabilities. By leveraging the multi-agent collaboration, we systematically decouple the satirical images into fine-grained semantic representations, as shown in Figure 3, enhancing the perception of visual information from local and global perspectives.

Prompt Construction and CoT Reasoning

The fine-grained semantic representations decoupled from images, including $\{LE_y, LE_b, GS_y, GS_b, D_l, D_g\}$, are employed to construct a CoT prompt to decompose the complex satire comprehension task into three subtasks, sequentially focusing on: local entity extraction, global semantic extraction, and satirical meaning inference, which facilitate step-by-step inference from local and global perspectives and guiding MLLM to focus on elementary details and explore the transformation or potential incongruous elements between the two scenes I_y and I_b when interpreting satirical images. By further analyzing the incongruities within their social or cultural contexts, MLLM is induced to explore the deep semantics of satire in conjunction with social issues. Herein, we denote the results of the subtasks as R_1 , R_2 , and R_3 . The overview of prompt is depicted in Figure 3.

Inference Optimization with Uncertainty Analysis

In complex scenes, current MLLMs show a propensity to misinterpret crucial image elements, overlook local entities, and fabricate non-existent objects, leading to hallucination issues (Huang et al. 2025) and unreliable outputs. Uncertainty refers to the level of confidence or the degree of unpredictability associated with the outputs of models and has been proven to have a significant effect on hallucination issues (Zhou et al. 2024). To mitigate hallucinations while comprehending satirical images, we propose an uncertainty analysis strategy during inference.

After the visual cascaded decoupling of satirical images, the Chain-of-Thought prompt directs the MLLM to orderly perform three subtasks in the process of satire comprehension inference. During this process, uncertainty analysis scores (UA scores) are computed for the results R_1 (about

Model	Size	Automatic Evaluation \uparrow					User Study \uparrow				
		BLEU	R-L	MT	BERT	AVE	Correct	Length	Complete	Faithful	AVE
MiniGPT4	7B	0.002	0.143	0.156	0.828	0.282	15.67	3.00	2.67	19.33	10.18
GPT4	-	0.003	0.151	0.219	0.852	0.306	58.00	31.67	37.00	45.33	43.00
Kosmos-2	-	0.011	0.202	0.197	0.867	0.319	15.33	5.33	7.67	9.00	9.33
Gemini	-	0.008	0.190	0.238	0.853	0.322	46.67	56.33	52.00	49.67	51.17
LLaVA	7B	0.011	0.180	0.225	0.859	0.319	25.67	19.67	23.00	26.33	23.67
LLaVA + ♣	7B	0.034	0.239	0.270	0.869	0.353	62.33	21.33	42.67	59.67	46.50
LLaVA	13B	0.014	0.197	0.226	0.860	0.324	28.33	21.00	27.67	29.33	26.58
LLaVA + ♣	13B	0.037	0.240	0.273	0.870	0.355	62.00	20.67	50.33	56.00	47.25
LLaVA-Next	7B	0.013	0.189	0.230	0.861	0.323	27.33	21.00	25.67	27.67	25.42
LLaVA-Next + ♣	7B	0.035	0.249	0.276	0.872	0.356	65.00	20.67	48.33	58.67	48.17
Qwen-VL	7B	0.014	0.177	0.219	0.854	0.316	31.00	21.33	29.67	34.33	29.08
Qwen-VL + ♣	7B	0.030	0.229	0.271	0.869	0.350	56.00	25.33	47.67	61.33	47.58
Qwen2.5-VL	7B	0.026	0.214	0.240	0.865	0.336	61.33	49.67	52.00	54.33	54.33
Qwen2.5-VL + ♣	7B	0.038	0.247	0.279	0.873	0.360	71.33	50.33	64.67	72.00	64.58

Table 1: Comparison of different baselines in satirical image comprehension. We conduct a user study and automatic evaluation of NLG metrics. The symbol “♣” represents SatireDecoder. “R-L”, “MT”, “BERT”, and “AVE” stand for ROUGE-L, METEOR, BERTScore, and the average scores computed across other metrics. User study is based on four criteria: correctness of the satire interpretation, appropriateness of the interpretation length, visual completeness interpretation, and faithfulness to the visual objects. Each criterion is evaluated by three users with a binary “yes” or “no” judgment. Baselines include MiniGPT4 (Zhu et al. 2023), GPT4 (OpenAI 2023), Kosmos-2 (Peng et al. 2023), Gemini (Team 2023), LLaVA (Liu et al. 2024b), LLaVA-Next (Liu et al. 2024a), Qwen-VL (Bai, Bai, and Yang 2023), Qwen2.5-VL (Bai and Chen 2025)

local entities) and R_2 (about global semantics) generated by the subtask-1 and subtask-2, against the standardized outputs LE_R_1 and GS_R_2 of Local Entities Extraction Agent LE and Global Semantics Extraction Agent GS , respectively. This procedure is repeated multiple times with varying model temperature settings of MLLM to minimize the UA scores of subtask-1 and subtask-2. Temperature is a parameter employed to regulate the creativity level (Zhu et al. 2024a) in text generation by language models. Given the logits Z_i for each candidate word, the corresponding probability distribution $P(i)$ is computed as:

$$P(i) = \frac{e^{z_i/Temp}}{\sum_j e^{z_j/Temp}} \quad (1)$$

where $Temp$ denotes model temperature, i represents the index of the target word under evaluation, and j corresponds to the index of all words in the vocabulary. Temperature modulates the probability distribution of the model, influencing the generation process by making the output content either more focused and deterministic or more random and diverse. Increasing temperature promotes greater diversity in generated content, revealing deeper comprehension and potentially hidden meanings. However, higher temperature also increases the risk of hallucinations and logical incoherence. Conversely, lower temperatures provide more stable and coherent outputs, reducing randomness but potentially overlooking subtle or latent implications. Our method reduces the model’s uncertainty regarding intermediate reasoning steps during complex reasoning tasks, thereby achieving the objective of controlling the reasoning path to obtain better responses for the final subtask-3 (about satire comprehension) and mitigate hallucinations inherent in multi-step inference processes.

Specifically, the results LE_R_1 and R_1 derived from the Local Entities Extraction Agent and MLLM subtask-1 are sets of local entity tags within satirical images. The uncertainty is quantified as the opposite number of the Jaccard similarity coefficient (Jaccard 1912), as follows:

$$U_1 = \min\left\{Temp\left(-\frac{|LE_R_1 \cap R_1|}{|LE_R_1 \cup R_1|}\right)\right\} \quad (2)$$

Furthermore, the results GS_R_2 and R_2 derived from the Global Semantics Agent and MLLM subtask-2 are captions of satirical images. The uncertainty is quantified as the opposite number of the BERTScore (Zhang et al. 2020), which leverages contextual embeddings from pre-trained language models (Devlin et al. 2019; Liu et al. 2019) to measure the semantic similarity between a candidate text and a reference text. The UA score of GS_R_2 and R_2 can be expressed as:

$$U_2 = \min\{Temp(-BERTScore(GS_R_2, R_2))\} \quad (3)$$

By controlling the model temperature hyperparameter to minimize uncertainty in the CoT reasoning, the result for subtask-3 exhibiting the least uncertainty is obtained as the final response for the satire comprehension task in our study.

Experimental Setup

Baseline. To ensure the consistency of the experiments, we follow (Nandy et al. 2024), utilizing the baselines including MiniGPT4 (Zhu et al. 2023), GPT4 (OpenAI 2023), Kosmos-2 (Peng et al. 2023), Gemini (Team 2023). Among them, MiniGPT4 performs worst due to the restricted leverage of visual information compared to text. Despite demonstrating notable cross-modal reasoning and visual grounding capabilities, both Gemini and Kosmos-2 face prominent

Model	Correct \uparrow	Length \uparrow	Complete \uparrow	Faithful \uparrow	AVE \uparrow	CHAIR _i \downarrow	CHAIR _s \downarrow
LLaVA+♣	62.33	21.33	42.67	59.67	46.50	36.53	41.02
LLaVA+♣ (w/o UA)	43.33	20.00	28.67	47.33	34.83	55.39	59.17
LLaVA-Next+♣	65.00	20.67	48.33	58.67	48.17	34.80	39.75
LLaVA-Next+♣ (w/o UA)	47.67	21.00	36.33	41.00	36.50	49.53	55.24
Qwen-VL+♣	56.00	25.33	47.67	61.33	47.58	39.83	49.01
Qwen-VL+♣ (w/o UA)	34.67	22.00	29.67	45.00	32.84	54.79	59.64
Qwen2.5-VL+♣	71.33	50.33	64.67	72.00	64.58	26.90	35.62
Qwen2.5-VL+♣ (w/o UA)	65.67	49.33	54.00	59.67	57.17	39.75	49.28

Table 2: Ablation study of the uncertainty analysis in SatireDecoder. To evaluate the effect of hallucination mitigation, the ablation experiment is based on the user study and CHAIR metrics from the object and sentence levels. The symbol “♣” represents SatireDecoder. “w/o” stands for “without”. “UA” represents uncertainty analysis during inference.

challenges in the global context analysis. We also select several SOTA backbones to explore and validate the effectiveness of SatireDecoder, including LLaVA (Liu et al. 2024b), LLaVA-NeXT (Liu et al. 2024a), Qwen-VL (Bai, Bai, and Yang 2023), and Qwen2.5-VL (Bai and Chen 2025). The inference optimization with uncertainty analysis is conducted with the temperature hyperparameters from 0.2 to 1.0.

Comparison of Baselines. Primarily, we perform the automatic evaluation to conduct a fair comparison among the baseline models and our proposed method, using the natural language generation (NLG) metrics, including BLEU (Papineni et al. 2002), ROUGE-L (Lin 2004), METEOR (Banerjee and Lavie 2005), BERTScore (Zhang et al. 2020), and an average score of the four NLG metrics.

As the result shown in Table 1, the baseline models MiniGPT4, GPT4, Kosmos-2, and Gemini exhibit disappointing performance. The average scores (normalized between 0 and 1) of automatic evaluation of baseline models are below 0.34, while the MLLM backbones equipped with SatireDecoder surpass the base models by approximately 4%. Furthermore, as the results indicate, our proposed SatireDecoder demonstrates a clear superiority in the NLG metrics, which assess the extent of n-gram matching, semantic correspondence, variation in vocabulary, syntactic patterns, and the completeness of essential information between the generated and reference texts, providing a comprehensive assessment of the satire comprehension quality.

User Study. To evaluate whether the satire comprehension more accurately aligns with human intentions, we sample 100 images from YesBut, conducting a user study of the generated content by three users with a binary “yes” or “no” judgment. The evaluation is based on four criteria: correctness of the satire interpretation, appropriateness of the interpretation length, visual completeness interpretation, and faithfulness to the visual objects. Each criterion is evaluated by three users with a binary “yes” or “no” judgment. As shown in Table 1, the baseline models encounter difficulties in accurately interpreting the satirical meanings within images, resulting in interpretations that lack precision in capturing the visual nuances. In contrast, our approach significantly surpasses the baseline models and backbones in the correctness of satire comprehension and the integrity and faithfulness of visual information representation by approximately 10% to 40%, effectively mitigating hallucinations in

satire comprehension. By incorporating SatireDecoder during inference, LLaVA achieves significant improvements in correctness, completeness, and faithfulness, enhancing the above metrics by approximately 37%, 20%, and 33%, respectively, with minimal computational overhead.

Ablation Study. To validate the effectiveness of uncertainty analysis in SatireDecoder, we conduct an ablation study. As shown in Table 2, the user study results indicate that uncertainty analysis significantly enhanced the performance of backbones in terms of the correctness of the satire comprehension, the visual completeness of the generated text, and the faithfulness to the visual objects. Specifically, the improvements are approximately 6% to 20% for correctness, about 10% to 16% for visual completeness, and roughly 12% to 18% for faithfulness.

Furthermore, to validate the effectiveness on mitigating hallucinations, we employ the CHAIR metric (Rohrbach et al. 2018) to measure the object hallucinations arising in the satire comprehension. CHAIR metric calculates the proportion of generated words that accurately correspond to the local objects in an image, as determined by the ground truth sentences and object segmentations. CHAIR is extended into two variants: $CHAIR_i$, which represents the fraction of object instances that are hallucinated, and $CHAIR_s$, which stands for the fraction of sentences that include a hallucinated object. Let H_o and Num_o denote hallucinated objects and all objects mentioned. And let H_s and Num_s stand for sentences with hallucinated objects and all sentences. CHAIR metrics are computed as:

$$CHAIR_i = \frac{|H_o|}{|Num_o|} \quad (4)$$

$$CHAIR_s = \frac{|H_s|}{|Num_s|} \quad (5)$$

As shown in Table 2, uncertainty analysis plays an important role in reducing the CHAIR metrics. The experimental results reveal that by restricting the inclusion of extraneous semantic entities in generated text during the inference phase, uncertainty analysis significantly reduces both object-level and sentence-level hallucinations, thereby enhancing the correctness of satirical image comprehension.

To validate the contribution of each component in the multi-agent system, we also conduct an ablation study. As

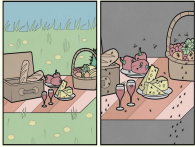


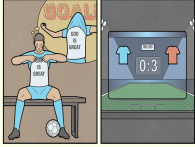





	LLaVA	LLaVA + SatireDecoder w/o “UA”	LLaVA + SatireDecoder with “UA”
(a) YES, BUT	 <p>The image is funny. Instead of the typical picnic elements like sand-wiches, fruit, and wine, the scene includes a suitcase, a beehive, and a cheese wheel. The presence of these unusual items creates a humorous and unexpected contrast to the typical picnic setting, making the image amusing and satirical.</p>	 <p>The image is humorous because it juxtaposes the "YES" half, which shows a man enjoying a picnic with a basket of food, and the "BUT" half, which shows a swarm of insects, specifically ants, attracted to the food. This creates a contrast between the pleasant and peaceful scene of the man enjoying his meal and the chaotic and disruptive presence of the insects.</p>	 <p>The image shows a garden setting with a basket of fruit, including apples and bananas, and a plate of food. However, the presence of insects or ants attracted to the food contradicts the idyllic garden atmosphere. This juxtaposition creates a humorous and ironic contrast, suggesting that even seemingly pleasant and natural settings can be tainted by the presence of pests.</p>
(b) YES, BUT	 <p>It suggests that the man is watching a soccer game on TV instead of actively participating in the sport. It highlights the contrast between the real-life experience of playing soccer and the more passive, sedentary experience of watching a game on TV.</p>	 <p>The image juxtaposes the seriousness of a football player's religious beliefs with the mundane and unrelated context of a football match. His t-shirt under jersey reads "GOD IS GREAT", but the BUT side shows the player's team losing the match. This contrast highlights the irony of the player's beliefs.</p>	 <p>The image juxtaposes the idea of a football player celebrating his goal by showing off his "GOD IS GREAT" t-shirt with the actual result of the match, where the player's team loses. The contra-diction between the player's optimism and the reality of the situation creates a comical effect.</p>
(c) YES, BUT	 <p>The image juxtaposes a soccer player's legs kicking a soccer ball with a pair of soccer cleats, which are placed on a pedestal. This creates a humorous and unexpected visual effect, as the cleats are elevated above the player's feet, making it seem as if they are performing a trick or defying gravity.</p>	 <p>The image is humorous because it juxtaposes the expensive, limited edition football boots with the man kicking a soccer ball. The contrast between the luxury item and the casual, everyday activity highlights the absurdity of spending a significant amount of money on sports equipment. This contrast makes the image satirical and amusing.</p>	 <p>The "YES" section features a pair of expensive-looking limited edition football boots, while the "BUT" side presents a man shooting a football wide outside a goalpost. This juxtaposition mocks the notion that expensive footwear can significantly improve one's performance in sports. The irony is that despite the costly boots, the outcome remains the same.</p>

Figure 4: Visualization of the ablation study. “UA” represents Uncertainty Analysis. The **red text** indicates the hallucinations and misinterpretations of satirical meaning. The **green text** highlights the objects and the correct satirical meanings newly captured after the application of SatireDecoder and uncertainty analysis.

Model	Corr. ↑	Len. ↑	Comp. ↑	Faith. ↑
LLaVA+♣	62.33	21.33	42.67	59.67
LLaVA+♣ (w/o LE)	50.33	20.33	37.67	38.33
LLaVA+♣ (w/o GS)	47.67	18.67	34.00	41.33
LLaVA+♣ (w/o DA)	54.00	19.67	38.33	42.67

Table 3: Ablation study for multi-agent system in visual cascaded decoupling. “LE”, “GS”, and “DA” stand for Local Entities Agent, Global Semantics Agent, and Discrepancy Analysis Agent, respectively. “Corr.”, “Len.”, “Comp.”, and “Faith.” stand for the four criteria in the user study.

shown in Table 3, each part of our multi-agent collaboration is beneficial for visual cascaded decoupling and satire comprehension.

Visualization. To visually demonstrate the effectiveness of our proposed method, we compare the results of SatireDecoder and baseline models, as shown in Figure 4. GPT4 and LLaVA either overlook or fabricate crucial local entities within images, resulting in misinterpretations of the deep semantics in satirical images. However, LLaVA equipped with SatireDecoder shows some improvement in detecting local entities and details, and further integration with uncertainty analysis strategy significantly enhances the satire comprehension capabilities, capturing more fine-grained details missed by baseline models. As shown in Figure 4 (a), LLaVA equipped SatireDecoder (without uncertainty analysis) exhibits object-level hallucination “the man enjoying

his meal”, which does not exist in the image. Under the effect of uncertainty analysis, the above hallucination has been killed. Moreover, as illustrated in Figure 4 (c), only LLaVA equipped with SatireDecoder and uncertainty analysis successfully captures the spatial position relationship of the soccer ball with the goalpost, indicating no goal, a crucial nuance that other models fail to recognize. More visual comparisons are presented in the Appendix submitted with Supplementary Materials. The improved ability allows MLLMs to better understand the relationship between local entities and the global context, leading to a more accurate comprehension of satirical meaning.

Conclusion

We propose a training-free approach, SatireDecoder, to enhance MLLMs’ perception and comprehension of satirical images, addressing the challenging multimodal satire comprehension task. Experiments demonstrate the effectiveness of multi-agent-based visual cascaded decoupling in boosting the perception of visual information, as well as the utility of CoT-based uncertainty analysis in mitigating hallucinations, making SatireDecoder a powerful and cost-effective approach for satirical image comprehension.

Acknowledgments

This project was funded by the National Natural Science Foundation of China (82090052).

References

- Bai, J.; Bai, S.; and Yang, S., et al. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. In *arxiv*.
- Bai, S.; and Chen, K., et al. 2025. Qwen2.5-VL Technical Report. In *arxiv*.
- Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL*.
- Bharti, S. K.; and Gupta, R. K., et al. 2022. Multimodal sarcasm detection: a deep learning approach. In *Wireless Commun. Mob. Comput.*
- Bi, J.; Wang, Y.; Chen, H.; Xiao, X.; Hecker, A.; Tresp, V.; and Ma, Y. 2025a. LLaVA Steering: Visual Instruction Tuning with 500x Fewer Parameters through Modality Linear Representation-Steering. In *ACL*.
- Bi, J.; Yan, D.; Wang, Y.; Huang, W.; Chen, H.; Wan, G.; Ye, M.; Xiao, X.; Schuetze, H.; Tresp, V.; et al. 2025b. CoTKinetics: A Theoretical Modeling Assessing LRM Reasoning Process. *arXiv preprint arXiv:2505.13408*.
- Bouazizi, M.; and Ohtsuki, T. 2015. Sarcasm detection in twitter: "all your products are incredibly amazing!!!"-are they really? In *GLOBECOM*.
- Bullier, J. 2001. Integrated model of visual processing. In *Brain Res. Rev.*
- Cai, Y., et al. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *ACL*.
- Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; and Poria, S. 2019. Towards multimodal sarcasm detection (an obviously perfect paper). In *ACL*.
- Chang, K.; Li, H.; Zhang, J.; and Wu, Y. 2024. NYK-MS: A Well-annotated Multi-modal Metaphor and Sarcasm Understanding Benchmark on Cartoon-Caption Dataset.
- Chen, J.; Jiang, Y.; Yang, D.; Li, M.; Wei, J.; Qian, Z.; and Zhang, L. 2024a. Can llms' tuning methods work in medical multimodal domain? In *MICCAI*.
- Chen, J.; Yang, D.; Jiang, Y.; Lei, Y.; and Zhang, L. 2024b. MISS: A generative pre-training and fine-tuning approach for med-VQA. In *ICANN*.
- Chen, J.; Yang, D.; Jiang, Y.; and Li, M., et al. 2024c. Efficiency in Focus: LayerNorm as a Catalyst for Fine-tuning Medical Visual Language Models. In *ACM MM*.
- Chen, J.; Yang, D.; Wu, T.; Jiang, Y.; Hou, X.; Li, M.; Wang, S.; Xiao, D.; Li, K.; and Zhang, L. 2024d. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*.
- Das, D.; and Clark, A. J. 2018. Sarcasm detection on facebook: A supervised learning approach. In *ICMI*.
- del Pilar Salas-Zárate, M.; Paredes-Valverde, M. A.; Rodríguez-García, M. Á.; Valencia-García, R.; and Alor-Hernández, G. 2017. Automatic detection of satire in Twitter: A psycholinguistic-based approach. In *Knowl.-Based Syst.*
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.
- Grill-Spector, K.; and Malach, R. 2004. The human visual cortex. In *Annu. Rev. Neurosci.*
- Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.
- Hyun, L.; Sung-Bin, K.; Han, S.; Yu, Y.; and Oh, T.-H. 2024. SMILE: Multimodal Dataset for Understanding Laughter in Video with Language Models. In *NAACL*.
- Jaccard, P. 1912. The distribution of the flora in the alpine zone. *New Phytol.*
- Jäncke, L.; Liem, F.; and Merillat, S. 2021. Are language skills related to structural features in Broca's and Wernicke's area? *Eur J Neurosci.*
- Jiang, B.; Xie, Y.; Wang, X.; Su, W. J.; Taylor, C. J.; and Mallick, T. 2024. Multi-modal and multi-agent systems meet rationality: A survey. In *ICML 2024 Workshop on LLMs and Cognition*.
- Jiang, Y.; Chen, J.; Yang, D.; and Li, M., et al. 2025a. CoMT: Chain-of-Medical-Thought Reduces Hallucination in Medical Report Generation. In *ICASSP*.
- Jiang, Y.; Li, J.; Liu, Y.; Yang, D.; Zhou, F.; and Kong, Q. 2025b. DanmakuTPPBench: A Multi-modal Benchmark for Temporal Point Process Modeling and Understanding. In *NeurIPS*.
- Joshi, A.; Bhattacharyya, P.; and Carman, M. J. 2017. Automatic sarcasm detection: A survey. *CSUR*.
- Joshi, A.; Sharma, V.; and Bhattacharyya, P. 2015. Harnessing context incongruity for sarcasm detection. In *ACL*.
- Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*.
- Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.
- Li, M.; Hou, X.; Liu, Z.; Yang, D.; Qian, Z.; Chen, J.; Wei, J.; Jiang, Y.; Xu, Q.; and Zhang, L. 2025. MCCD: Multi-Agent Collaboration-based Compositional Diffusion for Complex Text-to-Image Generation. In *CVPR*.
- Li, M.; Yang, D.; Liu, Y.; Wang, S.; Chen, J.; Wang, S.; Wei, J.; Jiang, Y.; Xu, Q.; Hou, X.; et al. 2024. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. In *NeurIPS*.
- Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *ACL*.
- Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. 2024b. Visual Instruction Tuning. In *NeurIPS*.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

- Mischler, G., et al. 2024. Contextual feature extraction hierarchies converge in large language models and the brain. *Nature Machine Intelligence*.
- Nandy, A.; Agarwal, Y.; Patwa, A.; Das, M. M.; Bansal, A.; Raj, A.; Goyal, P.; and Ganguly, N. 2024. YesBut: A High-Quality Annotated Multimodal Dataset for evaluating Satire Comprehension capability of Vision-Language Models. In *EMNLP*.
- OpenAI. 2023. GPT-4 Technical Report.
- Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*.
- Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World.
- Ptáček, T.; Habernal, I.; and Hong, J. 2014. Sarcasm detection on Czech and English Twitter. In *COLING*.
- Qin, L.; Huang, S.; Chen, Q.; Cai, C.; Zhang, Y.; Liang, B.; Che, W.; and Xu, R. 2023. MMSD2. 0: Towards a reliable multi-modal sarcasm detection system. In *ACL*.
- QwenTeam; et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object hallucination in image captioning. In *EMNLP*.
- Schifanella, R.; De Juan, P.; Tetreault, J.; and Cao, L. 2016. Detecting sarcasm in multimodal social platforms. In *ACM MM*.
- Team, G. 2023. Gemini: A Family of Highly Capable Multimodal Models.
- Thu, P. P.; and Aung, T. N. 2018. Implementation of emotional features on satire detection. In *Int. J. Netw. Distrib. Comput.*
- Wang, Y.; Bi, J.; Ma, Y.; and Pirk, S. 2025. ASCD: Attention-Steerable Contrastive Decoding for Reducing Hallucination in MLLM. *arXiv preprint arXiv:2506.14766*.
- Wang, Z.; Duan, J.; Cheng, L.; Zhang, Y.; Wang, Q.; Shi, X.; Xu, K.; Shen, H. T.; and Zhu, X. 2024. ConU: Conformal Uncertainty in Large Language Models with Correctness Coverage Guarantees. In *EMNLP*.
- Wu, C., et al. 2018. Thu_ngn at semeval-2018 task 3: Tweet irony detection with densely connected LSTM and multi-task learning. In *IWSM*.
- Xie, Y.; Zhu, Z.; Chen, X.; Chen, Z.; and Huang, Z. 2024. MoBA: Mixture of Bi-directional Adapter for Multi-modal Sarcasm Detection. In *ACM MM*.
- Xue, H.; Luo, X.; Hu, Z.; Zhang, X.; Xiang, X.; Dai, Y.; Liu, J.; Zhang, Z.; Li, M.; Yang, J.; Ma, F.; Wu, Z.; Yang, C.; Dai, Z.; and Yu, F. R. 2025. Human Motion Video Generation: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(11): 10709–10730.
- Yang, D.; Li, M.; Qu, L.; Yang, K.; Zhai, P.; Wang, S.; and Zhang, L. 2024a. Asynchronous Multimodal Video Sequence Fusion via Learning Modality-Exclusive and-Agnostic Representations. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Yang, D.; Liu, Y.; Huang, C.; Li, M.; Zhao, X.; Wang, Y.; Yang, K.; Wang, Y.; Zhai, P.; and Zhang, L. 2023. Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences. *Knowledge-Based Syst.*
- Yang, D.; Yang, K.; Kuang, H.; Chen, Z.; Wang, Y.; and Zhang, L. 2024b. Towards context-aware emotion recognition debiasing from a causal demystification perspective via de-confounded training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Yang, F.; Mukherjee, A.; and Dragut, E. 2017. Satirical news detection and analysis using attention mechanism and linguistic features. *EMNLP*.
- Yu, H.; Qi, Z.; Jang, L.; Salakhutdinov, R.; Morency, L.-P.; and Liang, P. P. 2024. MMoE: Enhancing Multimodal Models with Mixtures of Multimodal Interaction Experts. In *EMNLP*.
- Yue, T.; Shi, X.; Mao, R.; Hu, Z.; and Cambria, E. 2024. SarcNet: a multilingual multimodal sarcasm detection dataset. In *LREC-COLING*.
- Zeng, A.; Attarian, M.; Ichter, B.; Choromanski, K.; Wong, A.; Welker, S.; Tombari, F.; Purohit, A.; Ryoo, M.; Sindhvani, V.; Lee, J.; Vanhoucke, V.; and Florence, P. 2023. Socratic Models: Composing Zero-Shot Multimodal Reasoning with Language. In *ICLR*.
- Zhang, C.; Wan, Z.; Kan, Z.; Ma, M. Q.; Stepputtis, S.; Ramanan, D.; Salakhutdinov, R.; Morency, L.-P.; Sycara, K.; and Xie, Y. 2025. Self-Correcting Decoding with Generative Feedback for Mitigating Hallucinations in Large Vision-Language Models. In *ICLR*.
- Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *ICLR*.
- Zhang, Y.; Huang, X.; Ma, J.; Li, Z.; Luo, Z.; Xie, Y.; Qin, Y.; Luo, T.; Li, Y.; Liu, S.; et al. 2024. Recognize Anything: A Strong Image Tagging Model. In *CVPR*.
- Zhao, S.; Yuan, J.; Yang, G.; and Naseem, U. 2025. Can Pruning Improve Reasoning? Revisiting Long-CoT Compression with Capability in Mind for Better Reasoning. *arXiv preprint arXiv:2505.14582*.
- Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2024. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. In *ICLR*.
- Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models.
- Zhu, Y.; Li, J.; Li, G.; Zhao, Y.; Jin, Z.; and Mei, H. 2024a. Hot or Cold? Adaptive Temperature Sampling for Code Generation with Large Language Models. In *AAAI*.
- Zhu, Z.; Zhuang, X.; Zhang, Y.; Xu, D.; Hu, G.; Wu, X.; and Zheng, Y. 2024b. Tfcd: Towards multi-modal sarcasm detection via training-free counterfactual debiasing. In *IJCAI*.