

MemeScope: Probing and Causally Localizing Pragmatic Inference in LVLM Meme Understanding

Tongguan Wang^{1,2,3,4}, You Wu¹, Yuling Zhou¹, Siyuan Chen¹, Junkai Li¹, Feiyue Xue¹, Wei Chen¹, Hui Liu¹, Xiqiao Ba¹, Ying Sha^{1,2,3,4*},

¹College of Informatics, Huazhong Agricultural University

²Key Laboratory of Smart Farming for Agricultural Animals

³Engineering Research Center of Intelligent Technology for Agriculture, Ministry of Education

⁴Hubei Engineering Technology Research Center of Agricultural Big Data

wang_tg@webmail.hzau.edu.cn, shaying@mail.hzau.edu.cn

Abstract

Memes convey implicit pragmatic meanings through tightly coupled visual and textual cues, making their interpretation more challenging than that of conventional image-text pairs. Although large vision-language models (LVLMs) perform well on multimodal benchmarks, it remains unclear whether they infer non-literal mechanisms underlying memes, such as irony, wordplay, or rely mainly on superficial visual-textual correlations. To study this problem, we introduce MemeScope, a fine-grained benchmark of 1,781 memes annotated with eight pragmatic mechanisms and stage-wise explanations covering visual grounding, cross-modal alignment, and pragmatic inference. Building on MemeScope, we propose a unified interpretability framework that combines layer-wise probing with hidden-state patching to examine both the emergence and the causal contributions of meme-level pragmatic representations within three LVLMs. Our analysis shows that literal visual and textual cues are often captured in earlier layers, whereas pragmatic mechanism information emerges more strongly in higher layers and shows stronger patching effects in late-layer multimodal representations. We further identify a recurring pragmatic drift, where models ground meme content correctly but move toward plausible yet incorrect mechanisms during deeper decoding.¹

1 Introduction

Memes are a pervasive form of online communication that convey meaning through the interplay among visual content, overlaid text, and shared cultural knowledge (Wilkins, 1998). Unlike conventional image-text tasks, where the target semantics are often literal and explicitly grounded in the visual scene, meme understanding requires models to infer non-literal pragmatic meanings such

*Corresponding Author

¹Disclaimer: The samples presented by this paper may be considered offensive.

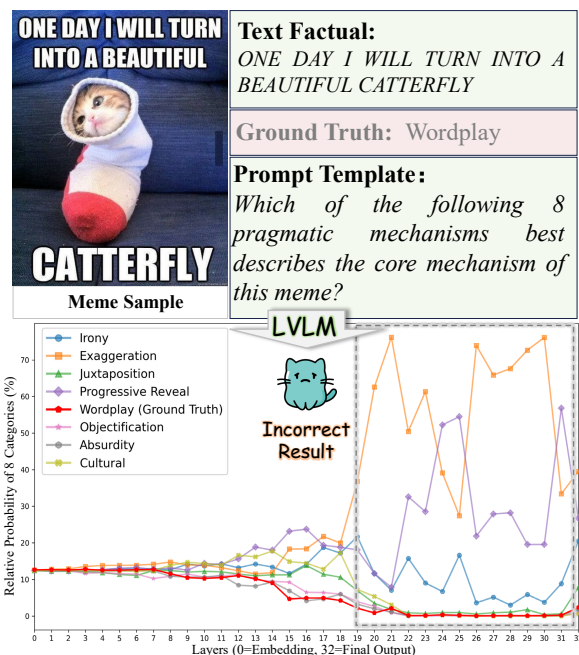


Figure 1: **Motivating example of pragmatic drift in LVLM meme understanding.** For a meme whose ground-truth pragmatic mechanism is *Wordplay*, we visualize the layer-wise probability assigned to each candidate mechanism by LLaVA-1.5-7B (Liu et al., 2023). Although the model initially maintains relatively diffuse predictions, deeper layers increasingly assign high probability to plausible but incorrect mechanisms such as *Exaggeration* and *Progressive Reveal*, while the correct *Wordplay* signal remains suppressed.

as irony, exaggeration, wordplay, absurdity, and cultural reference. This capability is essential for advancing multimodal language understanding and for building AI systems that can interpret complex human communication in realistic web environments, including tasks such as content moderation (Lai et al., 2022; Markov et al., 2023; Hacker et al., 2023; Wang et al., 2026b; Li et al., 2026) and harmful content detection (Zhang et al., 2025; Lu et al., 2024; Wang et al., 2026a,c; Fu et al., 2025; Hu et al., 2026; Wang et al., 2025).

Recent progress in LVLMs has led to strong performance on a wide range of multimodal benchmarks, including visual question answering (Sima et al., 2024; Qian et al., 2024), image captioning (Cheng et al., 2025; Kim et al., 2025), and instruction-following tasks (Liu et al., 2023, 2024). However, meme understanding remains substantially more difficult. Prior work has shown that meme-related tasks often require subtle multimodal reasoning and cannot be reliably solved by unimodal cues alone (Kiela et al., 2020). More recent datasets further highlight that models struggle to interpret visual metaphors and implicit meanings in memes, even when they perform well on standard vision-language tasks (Hwang and Shwartz, 2023). Nevertheless, most existing studies focus primarily on end-task performance, such as harmful meme detection or meme captioning, leaving an important question underexplored: ***When and where do LVLMs internally construct the pragmatic mechanisms required for meme understanding?***

To investigate this question, we perform a layer-wise diagnostic analysis of meme pragmatic reasoning in LLaVA-1.5-7B. Figure 1 shows a representative case where a cat wrapped in a sock is captioned with the invented word “catterfly”, whose intended mechanism is *Wordplay*. Although the humor depends on the portmanteau of “cat” and “butterfly” and its visual resemblance to a cocoon-like transformation, the model’s deeper layers shift probability mass toward plausible but incorrect mechanisms such as *Exaggeration* and *Progressive Reveal*, while the correct *Wordplay* signal remains suppressed. We use *pragmatic drift* to describe this failure pattern: the model may partially ground meme content, yet its later representations move toward an incorrect interpretation of the meme’s communicative intent. This suggests that meme understanding failures cannot be fully explained by poor visual recognition or OCR errors alone, but require analyzing how non-literal pragmatic cues emerge and drift inside LVLMs. However, existing meme datasets typically focus on task-level labels such as harmfulness, sentiment, or caption-level interpretation, while interpretability methods for vision-language models remain largely correlational. Thus, systematically studying LVLm meme reasoning requires both a fine-grained diagnostic benchmark and a causal framework for localizing pragmatic inference across layers and representations for this underexplored setting.

To address these challenges, we introduce

MemeScope, a fine-grained benchmark for pragmatic inference in meme understanding. MemeScope contains 1,781 memes annotated with eight pragmatic mechanisms: *Irony*, *Exaggeration*, *Juxtaposition*, *Progressive Reveal*, *Wordplay*, *Objectification*, *Absurdity*, and *Cultural Reference*. Beyond mechanism labels, each meme is annotated according to a three-stage decomposition of meme comprehension: *visual grounding*, which captures literal visual entities and scenes; *cross-modal alignment*, which captures how visual and textual cues interact or conflict; and *pragmatic inference*, which captures the implicit mechanism, background knowledge, and communicative intent.

Building on MemeScope, we propose a unified interpretability framework that combines layer-wise probing with hidden-state patching to analyze how LVLMs construct meme-level pragmatic meanings. Probing identifies where visual grounding, cross-modal alignment, and pragmatic inference become decodable from internal representations, while hidden-state patching causally tests whether factual representations can restore the correct pragmatic prediction under counterfactual inputs. Across representative LVLMs, our analysis shows that literal visual and textual cues tend to emerge earlier, cross-modal alignment becomes stronger in middle layers, and pragmatic mechanism information is more strongly encoded and causally concentrated in late-layer multimodal representations. We further identify recurring *pragmatic drift*, where models partially ground meme content but shift toward plausible yet incorrect mechanisms during deeper inference, revealing a persistent gap between multimodal perception and pragmatic reasoning.

In summary, our contributions are as follows:

- We introduce **MemeScope**, a fine-grained benchmark of 1,781 memes with 8 pragmatic mechanism labels and stage-wise explanatory annotations covering visual grounding, cross-modal alignment, and pragmatic inference.
- We develop a unified interpretability framework that combines layer-wise probing with hidden-state patching to study both the emergence and causal contributions of pragmatic representations inside LVLMs.
- Through extensive experiments, we identify pragmatic drift in LVLm meme understanding and provide patching-based evidence that

meme-level pragmatic inference relies heavily on higher-layer multimodal representations.

2 Related Work

2.1 Multimodal Meme Understanding

Multimodal meme understanding has evolved from task-specific classification toward richer semantic interpretation (Jha et al., 2024). As summarized in Table 1, existing datasets cover tasks ranging from harmful or offensive content classification (Suryawanshi et al., 2020; Kiela et al., 2020; Chandra et al., 2021; Pramanick et al., 2021; Fersini et al., 2022; Liu et al., 2022; Sharma et al., 2022; Lin et al., 2026) to explanation-oriented generation (Hee et al., 2023; Sharma et al., 2023; Hwang and Schwartz, 2023; Park et al., 2024; Zhao et al., 2025). Early benchmarks such as HatefulMemes (Kiela et al., 2020) and MultiOFF (Suryawanshi et al., 2020) focus on detecting hateful or offensive intent, establishing memes as a testbed for evaluating visual-textual integration. More recent datasets, such as MemeCap (Hwang and Schwartz, 2023) and MemeReaCon (Zhao et al., 2025), emphasize interpretation, captioning, reasoning, and contextual information. These resources show that meme understanding often requires background knowledge, figurative reasoning, and cross-modal semantic integration. However, most existing datasets are still organized around end-task prediction or free-form generation. They generally lack fine-grained pragmatic mechanism labels and are not designed to support mechanistic probing of how LVLMs internally construct the meaning of memes.

To address this gap, MemeScope complements prior meme benchmarks by shifting the focus from outcome-oriented evaluation to stage-aware analysis of multimodal pragmatic reasoning. Instead of merely asking whether a model produces the correct final label or explanation, MemeScope annotates each meme with fine-grained pragmatic mechanisms and stage-wise explanations that cover visual grounding, cross-modal alignment, and pragmatic inference. This design allows us to diagnose whether an LVLM fails because it cannot recognize relevant visual entities, cannot align them with textual cues, or cannot infer the intended non-literal mechanism. Moreover, by explicitly organizing meme annotations around these stages, MemeScope provides a basis for layer-wise probing and causal localization, enabling analysis of

where pragmatic inference emerges and how it is mediated inside LVLMs.

2.2 Mechanistic Interpretability in LVLMs

Mechanistic interpretability aims to move beyond end-to-end evaluation by examining the internal computations that give rise to model behavior. In language models, probing, causal mediation analysis, activation patching, and model editing have been widely used to identify where specific information is represented and whether it causally affects model outputs (Hewitt and Liang, 2019; Vig et al., 2020; Meng et al., 2022; Conmy et al., 2023). Probing methods can reveal whether information is decodable from hidden states, but they remain correlational and do not by themselves establish whether the model actually uses that information (Hewitt and Liang, 2019; Belinkov, 2022). In contrast, causal methods such as mediation analysis and activation patching provide stronger evidence by intervening on internal representations and measuring their effects on model predictions (Vig et al., 2020; Meng et al., 2022). Compared with text-only language models, the mechanistic analysis of LVLMs is still emerging. Existing work has begun to study visual token processing, visual-textual alignment, compositional visual reasoning, counting, and object-centric reasoning in multimodal models (Aravindan et al., 2025; Neo et al., 2025; Liu et al., 2025; Che et al., 2026). However, most existing analyses focus on relatively literal visual or compositional capabilities. The internal mechanisms underlying abstract social meaning, figurative interpretation, and pragmatic inference remain much less understood, even though such abilities are central to meme understanding. In memes, the intended meaning may arise from image-text incongruity, cultural reference, visual metaphor, or wordplay, requiring models to infer meanings beyond literal visual recognition and text reading (Grice, 1975; Sperber and Wilson, 1986).

We extend mechanistic interpretability to non-literal multimodal reasoning by decomposing meme comprehension into an operational three-stage framework: visual grounding, cross-modal alignment, and pragmatic inference. Leveraging MemeScope’s stage-wise annotations and counterfactual inputs, we integrate layer-wise probing with hidden-state patching. This unified approach distinguishes representations that merely correlate with pragmatic mechanisms from those that causally drive the final decision, enabling us to precisely

Dataset	Task Type	Implicit Pragmatics	Mechanistic Probing	Size
MultiOFF (Suryawanshi et al., 2020)	Classification	✗	✗	743
HatefulMemes (Kiela et al., 2020)	Classification	✗	✗	10k
Jewtocracy (Chandra et al., 2021)	Classification	✗	✗	6,611
HarMeme (Pramanick et al., 2021)	Classification	✗	✗	3,544
MAMI (Fersini et al., 2022)	Classification	✗	✗	15k
FigMemes (Liu et al., 2022)	Classification	✗	✗	5,141
HVVMemes (Sharma et al., 2022)	Classification	✗	✗	7k
GOAT (Lin et al., 2026)	Classification	✗	✗	6,626
HatReD (Hee et al., 2023)	Generation	✗	✗	3,228
ExHV V (Sharma et al., 2023)	Generation	✗	✗	4,680
MemeCap (Hwang and Shwartz, 2023)	Generation	✓	✗	6,387
MemeIntent (Park et al., 2024)	Generation	✓	✗	950
MemeReaCon (Zhao et al., 2025)	Classification & Generation	✓	✗	1,565
MemeScope (Ours)	Interpretability	✓	✓	1,781

Table 1: Comparison of MemeScope with representative multimodal meme datasets. MemeScope differs from prior datasets by providing fine-grained implicit pragmatic annotations and supporting mechanistic probing of LVLM meme understanding.

Category	Count	Percentage
Irony	401	22.5%
Exaggeration	273	15.3%
Juxtaposition	286	16.1%
Progressive Reveal	70	3.9%
Wordplay	187	10.5%
Objectification	216	12.1%
Absurdity	119	6.7%
Cultural Reference	229	12.9%
Total	1,781	100.0%

Table 2: Distribution of pragmatic mechanism labels in MemeScope.

localize where LVLMs construct, distort, or lose implicit meanings across internal layers.

3 MemeScope Dataset

To support fine-grained evaluation and mechanistic analysis of LVLM meme understanding, we construct **MemeScope**², which annotates both pragmatic mechanisms and stage-wise explanations. This allows us to diagnose whether LVLM failures arise from visual grounding, cross-modal alignment, or pragmatic inference.

3.1 Data Collection and Selection

MemeScope is built from four representative multimodal meme datasets: MultiOFF (Suryawanshi et al., 2020), ExHV V (Sharma et al., 2023), MemeCap (Hwang and Shwartz, 2023), and MemeReaCon (Zhao et al., 2025). We retain only memes that require non-literal pragmatic reasoning beyond literal image or text cues, such as image-text in-

congruity, cultural reference, wordplay, figurative expression, or commonsense violation. Further construction details are provided in Appendix A.1.

3.2 Annotation Schema

MemeScope decomposes meme comprehension into three stages: *visual grounding*, which describes literal entities and actions; *cross-modal alignment*, which captures how text relates to visual elements; and *pragmatic inference*, which identifies the non-literal mechanism and communicative intent. Each meme is assigned one of eight labels: *Irony*, *Exaggeration*, *Juxtaposition*, *Progressive Reveal*, *Wordplay*, *Objectification*, *Absurdity*, and *Cultural Reference*. Their definitions are provided in Appendix A.2; when multiple mechanisms are plausible, annotators select the dominant one that best explains the meme’s core communicative effect. Detailed annotations are shown in Appendix A.3.

3.3 Annotation Quality Control

We adopt a data annotation process that combines the assistance of a multimodal large language model and manual review. We employ GPT-5.4³ first to generate a structured draft, and then hire three doctoral candidates to conduct the review. The reviewers would check the visual descriptions, the mapping between text and images, pragmatic labels, background knowledge, and intention descriptions. In cases of disagreement, discussions would be held to resolve the issues. Further details of the annotation quality control process are provided in Appendix A.4.

²The data can be accessed [here](#).

³<https://openai.com/index/introducing-gpt-5-4/>

3.4 Dataset Statistics

Ultimately, the MemeScope dataset contains 1,781 annotated memes across eight pragmatic mechanisms. The full distribution is shown in Table 2.

4 Method

Figure 2 illustrates our framework for interpreting LVLM meme understanding. Given a meme input, we analyze not only whether the model predicts the correct pragmatic mechanism, but also when this information emerges and which internal states causally affect the decision. Our framework combines *layer-wise probing*, which identifies where visual grounding, cross-modal alignment, and pragmatic inference become decodable, with *hidden-state patching*, which localizes the layers and token groups that mediate the final pragmatic prediction.

4.1 Task Formulation

Let $\mathcal{M}$ denote an LVLM with L transformer layers. For each meme input

$$x = (I, T, P), \quad (1)$$

where I is the meme image, T is the extracted or provided meme text, and P is the task prompt specifying the candidate pragmatic mechanisms, $\mathcal{M}$ predicts a pragmatic mechanism

$$y_{\text{prag}} = \mathcal{M}(x), \quad y_{\text{prag}} \in \mathcal{Y}. \quad (2)$$

Here, $\mathcal{Y}$ contains eight categories: *Irony*, *Exaggeration*, *Juxtaposition*, *Progressive Reveal*, *Wordplay*, *Objectification*, *Absurdity*, and *Cultural Reference*.

4.2 State Extraction

The visual encoder converts the image I into visual features, which are mapped into the language model embedding space as visual tokens X_v . The meme text and prompt are tokenized into text and prompt tokens, denoted as X_t and X_p . The resulting input sequence is

$$X = [X_p; X_t; X_v], \quad (3)$$

where the ordering follows the corresponding model-specific input template.

During the forward pass, the transformer backbone produces hidden states at each layer.

$$H^{(l)} = [h_1^{(l)}, h_2^{(l)}, \dots, h_n^{(l)}], \quad (4)$$

Let $H^{(l)}$ denote the hidden states of all input tokens at layer l , where $h_i^{(l)} \in \mathbb{R}^d$. We use pooled representations over selected token groups for analysis.

Unless otherwise specified, the default representation for layer-wise probing is the hidden state at the final input position, denoted as $h_x^{(l)}$. For token-group causal analysis, we additionally consider visual tokens, text tokens, prompt tokens, and fused multimodal states.

4.3 Layer-wise Probing for Semantic Emergence

Layer-wise probing is used to identify at which stages of meme understanding the model’s internal representations become decodable. We consider three stages aligned with the MemeScope annotation schema: visual grounding, cross-modal alignment, and pragmatic inference. For each layer l and each stage $s \in \{\text{ground}, \text{align}, \text{prag}\}$, we train a lightweight linear probe on top of the frozen LVLM representation.

Given the layer- l representation $h_x^{(l)}$, the probe predicts the stage-specific label as

$$p_s^{(l)}(y | x) = \text{softmax} \left(W_s^{(l)} h_x^{(l)} + b_s^{(l)} \right), \quad (5)$$

where $W_s^{(l)}$ and $b_s^{(l)}$ are the parameters of the probe for stage s at layer l . The LVLM parameters remain frozen throughout probing. We optimize only the probe parameters using cross-entropy loss:

$$\mathcal{L}_{\text{probe}}^{(l,s)} = -\mathbb{E}_{(x,y_s)} \log p_s^{(l)}(y_s | x), \quad (6)$$

where y_s is the MemeScope annotation for stage s .

High probing performance indicates that stage-specific information is linearly decodable at a given layer, but it does not show whether the model actually uses this information for prediction. We therefore use probing to characterize layer-wise emergence, and then apply hidden-state patching to test whether these representations causally affect the final pragmatic decision.

4.4 Causal Intervention via Hidden-State Patching

While probing identifies where information is encoded, it does not reveal whether the model uses it for prediction. To test causal contribution, we perform hidden-state patching using factual input x_f (preserving the original pragmatic cue) and counterfactual input x_c (replacing the meme text with a literal visual description). Let $h_f^{(l)}$ and $h_c^{(l)}$ denote the hidden states at layer l for the factual and counterfactual runs.

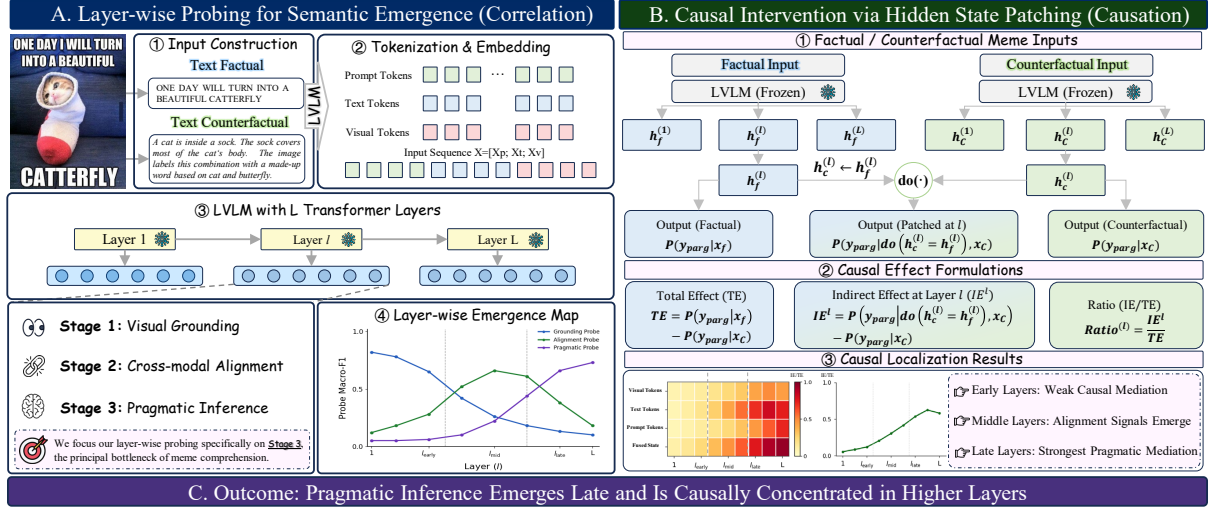


Figure 2: **Overview of our interpretability framework.** Panel A performs layer-wise probing to analyze the emergence of visual grounding, cross-modal alignment, and pragmatic inference across transformer layers. Panel B conducts causal intervention via hidden-state patching, where factual hidden states are injected into counterfactual runs to measure their causal contribution to the final pragmatic prediction.

We first compute the model’s probability of the target pragmatic mechanism under both unmodified inputs. The total effect of the factual pragmatic cue is defined as

$$TE = P(y_{\text{prag}} | x_f) - P(y_{\text{prag}} | x_c). \quad (7)$$

A larger TE indicates that removing the pragmatic cue substantially reduces the model’s confidence in the correct mechanism.

To measure the causal contribution of layer l , we run the model on the counterfactual input x_c and replace the last token’s hidden state with the corresponding factual hidden state (e.g., the final text token or [CLS]/pooled token in the fused multimodal representation):

$$h_{c,\text{last}}^{(l)} \leftarrow h_{f,\text{last}}^{(l)}. \quad (8)$$

This intervention is denoted using the $do(\cdot)$ operator. The indirect effect mediated by the last token at layer l is

$$IE^{(l)} = P(y_{\text{prag}} | do(h_{c,\text{last}}^{(l)} = h_{f,\text{last}}^{(l)}), x_c) - P(y_{\text{prag}} | x_c). \quad (9)$$

Intuitively, $IE^{(l)}$ measures how much the factual representation at the last token of layer l restores the correct pragmatic prediction when inserted into the counterfactual run.

To normalize the mediated effect across examples, we compute the effect ratio

$$\text{Ratio}^{(l)} = \frac{IE^{(l)}}{TE}. \quad (10)$$

We compute this quantity across layers to obtain a layer-wise causal localization map. In addition, we perform token-group patching by applying the same intervention to different subsets of hidden states, including visual tokens, text tokens, prompt tokens, and fused multimodal states. This allows us to identify not only which layers are causally important, but also which modality-specific or fused representations mediate pragmatic inference.

4.5 Interpretation of the Framework

The two components of our framework play distinct roles. Layer-wise probing identifies when stage-specific information becomes decodable, while hidden-state patching tests whether it causally affects the final pragmatic prediction. Together, they distinguish representations that are merely encoded from those that actually mediate meme-level pragmatic inference. This is crucial for analyzing LVLs that may ground visual or textual cues correctly but still drift toward an incorrect pragmatic mechanism in later layers.

5 Experimental Setup

We evaluate three representative open-weight LVLs: **LLaVA-1.5-7B** (Liu et al., 2023), **Qwen2.5-VL-7B** (Bai et al., 2025b), and **Qwen3-VL-8B** (Bai et al., 2025a). All LVL parameters are kept frozen throughout evaluation, probing, and patching. We use stratified 5-fold cross-validation on MemeScope, with 80% of samples used for

Model	Input	Acc.	Macro-F1	Ground. F1	Align. F1
LLaVA-1.5-7B	Full	47.3%	41.8%	66.2%	51.4%
	OCR/Text	39.5%	34.2%	30.8%	37.5%
	Image	35.8%	29.6%	59.4%	33.2%
Qwen2.5-VL-7B	Full	53.8%	48.7%	71.6%	57.9%
	OCR/Text	44.6%	39.7%	33.4%	42.8%
	Image	40.7%	35.1%	64.8%	38.6%
Qwen3-VL-8B	Full	58.2%	53.4%	75.8%	62.3%
	OCR/Text	48.9%	44.1%	36.5%	46.8%
	Image	44.5%	38.9%	68.7%	42.7%

Table 3: Main results on MemeScope. Acc. and Macro-F1 evaluate 8-way pragmatic mechanism classification. Stage-wise scores are reported as Macro-F1.

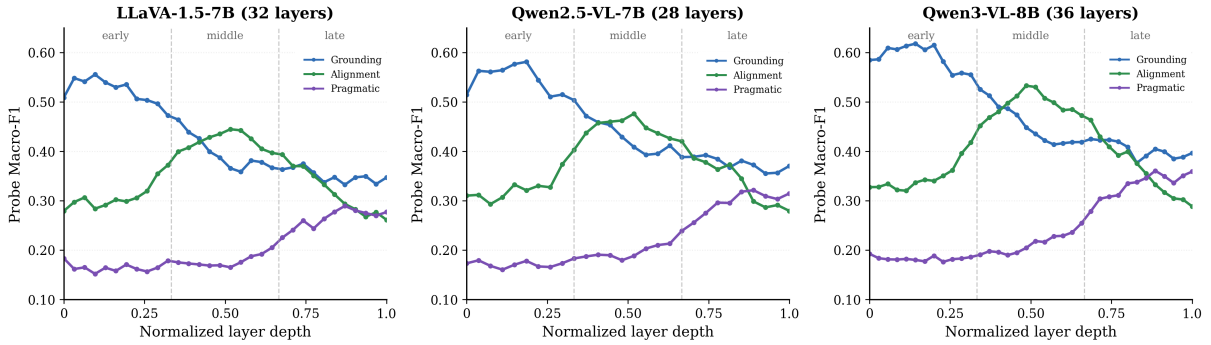


Figure 3: Layer-wise probing performance over normalized layer depth for LVLMs with different numbers of transformer layers. Grounding signals are strongest in early layers, alignment peaks around middle layers, and pragmatic signals increase mainly in later layers. All scores are reported as Macro-F1.

training or development and 20% held out for testing in each fold. Since the label distribution is naturally imbalanced, we use Accuracy and Macro-F1 as the primary metric. Direct LVLM evaluation is performed without fine-tuning; the development split is used only for prompt calibration. Layer-wise probes are trained on the training split and evaluated on the held-out split, whereas hidden-state patching is performed only on the held-out examples. The main experimental setup is summarized in Appendix B and Appendix C.

6 Results and Analysis

6.1 RQ1: Where Do LVLMs Fail in Meme Comprehension?

Table 3 reports the main results on MemeScope. Full multimodal input consistently outperforms OCR/Text-only and Image-only settings, showing that pragmatic meme understanding benefits from joint visual-textual modeling. Still, the strongest model remains far from solving the 8-way non-literal mechanism classification task.

The stage-wise results reveal a clear bottleneck: performance is highest on visual grounding, lower

on cross-modal alignment, and lowest on pragmatic inference. This suggests that LVLMs can often recognize literal meme content, but still struggle to infer the intended non-literal mechanism.

6.2 RQ2: How Does Pragmatic Meaning Emerge Across Layers?

Figure 3 reports layer-wise probing results over normalized layer depth, allowing comparison across LVLMs with different numbers of transformer layers. Across models, grounding information is strongest in early layers, alignment information peaks around middle layers, and pragmatic mechanism information increases mainly in later layers.

Pragmatic probing remains consistently lower than grounding and alignment, indicating that meme-level pragmatic mechanisms are harder to linearly decode than literal visual evidence or cross-modal correspondences. This supports the three-stage design of MemeScope and motivates the causal analysis in RQ3.

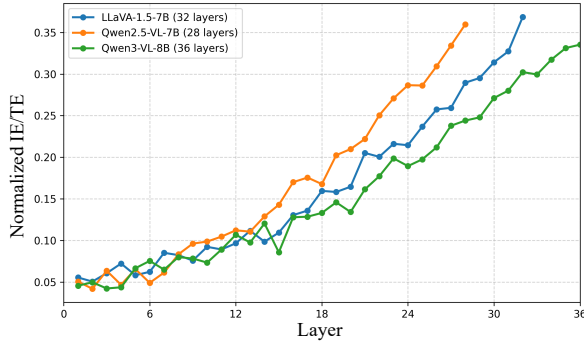


Figure 4: Layer-wise causal mediation of pragmatic inference measured by normalized indirect effect ratio (IE/TE).

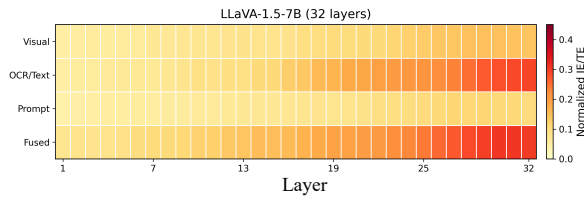


Figure 5: Layer $\times$ token-group causal effect heatmap for LLaVA-1.5-7B.

6.3 RQ3: Which Internal Representations Causally Mediate Pragmatic Inference?

We use hidden-state patching to test whether late-layer pragmatic signals influence LVLM predictions under controlled interventions. Figure 4 shows that patching effects are weak in early layers, rise through middle layers, and peak in late layers, suggesting a stronger patching-based causal contribution from higher-layer representations.

Figure 5 further breaks down causal contributions across token groups for LLaVA-1.5-7B. Fused multimodal states and OCR/Text tokens mediate the strongest effects, while visual and prompt tokens contribute less. The strongest effects occur primarily in middle-to-late and late layers, showing that meme-level pragmatic inference depends mainly on late-stage multimodal integration rather than visual grounding alone. Additional token-group analyses are provided in Appendix E.

6.4 RQ4: Do LVLMs Over-rely on Textual Cues?

Figure 6 shows that full input achieves the highest Macro-F1, confirming that joint visual-textual modeling benefits meme pragmatic understanding. Text-only input consistently outperforms Image-only input, highlighting the importance of textual cues. Counterfactual text x_c is created by neutral-

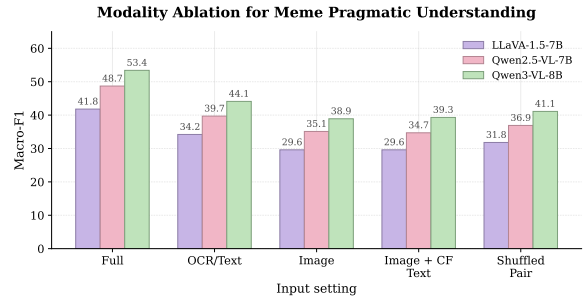


Figure 6: Modality ablation results measured by Macro-F1. Full multimodal input performs best, OCR/Text-only input outperforms Image-only, and counterfactual or shuffled inputs reduce performance, showing that LVLMs partially rely on textual cues.

izing or replacing only the core pragmatic trigger while preserving the sentence structure and narrative style. Shuffled image-text pairs reduce performance by breaking cross-modal alignment. Macro-F1 remains moderate, indicating that LVLMs partially rely on textual cues without fully verifying that the text is grounded in the image. This aligns with the inherent difficulty of 8-way non-literal pragmatic mechanism classification.

6.5 Qualitative Analysis: Pragmatic Drift in Meme Understanding

Figure 1 illustrates a representative pragmatic drift case. The catterfly meme requires identifying *Word-play*, since “catterfly” blends “cat” and “butterfly” while the sock-covered cat evokes a cocoon-like transformation. LLaVA-1.5-7B captures several local cues, but its later-layer predictions shift toward *Exaggeration* and *Progressive Reveal*, leaving the correct mechanism suppressed.

This failure is not simply an OCR or visual recognition error. The model partially grounds the meme content but fails to preserve the fine-grained relation that defines the intended pragmatic mechanism. This qualitative pattern aligns with RQ2 and RQ3: pragmatic information emerges late, remains weaker than grounding and alignment, and is mediated by late-layer fused and OCR/Text representations. More cases are shown in Appendix G.

7 Conclusion

In this paper, we introduce MemeScope, a fine-grained benchmark for LVLM meme understanding. Combining stage-wise annotations with probing and hidden-state patching, we show that pragmatic signals emerge later than grounding and

alignment, with stronger patching effects in late-layer fused multimodal and OCR/Text representations. We also identify pragmatic drift, where models partially ground meme content yet shift toward plausible but incorrect mechanisms. These findings highlight meme understanding as a testbed for studying the gap between multimodal perception and non-literal pragmatic reasoning in LVLMs.

Limitations

MemeScope is designed for fine-grained diagnosis rather than large-scale pretraining, and its size is therefore smaller than web-scale meme collections. Although all annotations are human-verified, meme interpretation can remain ambiguous, especially for culturally specific, temporally sensitive, or cross-lingual cases. In addition, our experiments focus on three representative open-weight LVLMs because our probing and hidden-state patching analyses require access to internal layer-wise representations. This design enables mechanistic analysis, but it does not cover closed-source frontier LVLMs whose hidden states are inaccessible. Extending MemeScope to larger and more culturally diverse meme collections, as well as evaluating closed-source LVLMs through prompting-only protocols, remains important future work.

Ethics Statement

MemeScope is intended for research on multimodal pragmatic reasoning and LVLM interpretability. Since memes may contain offensive, political, or culturally sensitive content, our annotations describe pragmatic mechanisms and communicative intent without endorsing the underlying messages. The dataset is constructed through LLM-assisted annotation followed by full human verification to reduce hallucinations and misinterpretations. Upon acceptance, we will release annotations, prompts, evaluation scripts, and dataset metadata for research use, subject to the redistribution constraints of the original meme images. Our findings also caution against using current LVLMs for fully automated moderation of non-literal meme content, as they may misinterpret pragmatic mechanisms and drift toward plausible but incorrect meanings. We do not attempt to identify individuals appearing in meme images, and all released metadata and annotations are intended for research use only. Original images remain subject to the access and redistribution terms of their source datasets.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 62272188). Thanks to the anonymous reviewers for their hard work!

References

- Ashwath Vaithinathan Aravindan, Abha Jha, and Mihir Kulkarni. 2025. Do vlms have bad eyes? diagnosing compositional failures via mechanistic interpretability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 715–723.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Ke qin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Qidong Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, and 46 others. 2025a. [Qwen3-vl technical report](#). *ArXiv*, abs/2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *CoRR*, abs/2502.13923.
- Yonatan Belinkov. 2022. [Probing classifiers: Promises, shortcomings, and advances](#). *Computational Linguistics*, 48(1):207–219.
- Mohit Chandra, Dheeraj Pailla, Himanshu Bhatia, Aadilmehdi Sanchawala, Manish Gupta, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021. [“subverting the jewtocracy”: Online antisemitism detection using multimodal deep learning](#). In *Proceedings of the 13th ACM Web Science Conference 2021*, WebSci ’21, page 148–157, New York, NY, USA. Association for Computing Machinery.
- Liwei Che, Zhiyu Xue, Yihao Quan, Benlin Liu, Zeru Shi, Michelle Hurst, Jacob Feldman, Ruixiang Tang, Ranjay Krishna, and Vladimir Pavlovic. 2026. [Counting circuits: Mechanistic interpretability of visual reasoning in large vision-language models](#). *CoRR*, abs/2603.18523.
- Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. 2025. [CapArena: Benchmarking and analyzing detailed image captioning in the LLM era](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14077–14094, Vienna, Austria. Association for Computational Linguistics.
- Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso.

2023. [Towards automated circuit discovery for mechanistic interpretability](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 16318–16352. Curran Associates, Inc.
- Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. [SemEval-2022 task 5: Multimedia automatic misogyny identification](#). In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.
- Shichao Fu, Tongguan Wang, and Ying Sha. 2025. [GEETI: graph embedding-enhanced textual inversion for chinese harmful meme detection and identification](#). In *Advanced Intelligent Computing Technology and Applications - 21st International Conference, ICIC 2025, Ningbo, China, July 26-29, 2025, Proceedings, Part X*, volume 15851 of *Lecture Notes in Computer Science*, pages 307–318. Springer.
- Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.
- Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. [Regulating chatgpt and other large generative ai models](#). In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23*, page 1112–1123, New York, NY, USA. Association for Computing Machinery.
- 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.
- John Hewitt and Percy Liang. 2019. [Designing and interpreting probes with control tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.
- Jiawei Hu, Yifan Lyu, Yifan Chen, Kuntao Li, Yun Xue, and Jinghua Liang. 2026. [Multi-view dynamic perception framework for chinese harmful meme detection](#). *Information Processing & Management*, 63(4):104602.
- EunJeong Hwang and Vered Shwartz. 2023. [MemeCap: A dataset for captioning and interpreting memes](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.
- Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024. [Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 930–943, St. Julian's, Malta. Association for Computational Linguistics.
- 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*, volume 33, pages 2611–2624. Curran Associates, Inc.
- Taewhan Kim, Soeun Lee, Si-Woo Kim, and Dong-Jin Kim. 2025. [Vipcap: Retrieval text-based visual prompts for lightweight image captioning](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):4320–4328.
- Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. [Human-ai collaboration via conditional delegation: A case study of content moderation](#). In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22*, New York, NY, USA. Association for Computing Machinery.
- Ruobing Li, Yifan Feng, Lin Shen, Liuxian Ma, Haojie Zhang, Kun Qian, Bin Hu, Yoshiharu Yamamoto, and Björn W. Schuller. 2026. [Fedvcpl-diff: A federated convolutional prototype learning framework with a diffusion model for speech emotion recognition](#). *Information Fusion*, 127:103745.
- Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. 2026. [Goat-bench: Safety insights to large multimodal models through meme-based social abuse](#). *ACM Transactions on Intelligent Systems and Technology*, 17(4):1–25.
- Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022. [FigMemes: A dataset for figurative language identification in politically-opinionated memes](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. [Improved baselines with visual instruction tuning](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.
- Yiming Liu, Yuhui Zhang, and Serena Yeung-Levy. 2025. [Mechanistic interpretability meets vision language models: Insights and limitations](#). In *The Fourth Blogpost Track at ICLR 2025*.

- Junyu Lu, Bo Xu, Xiaokun Zhang, Hongbo Wang, Hao-hao Zhu, Dongyu Zhang, Liang Yang, and Hongfei Lin. 2024. [Towards comprehensive detection of chinese harmful memes](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 13302–13320. Curran Associates, Inc.
- Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. 2023. [A holistic approach to undesired content detection in the real world](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12):15009–15018.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in gpt](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.
- Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2025. [Towards interpreting visual information processing in vision-language models](#). In *International Conference on Learning Representations*, volume 2025, pages 57172–57189.
- Jeongsik Park, Khoi P. N. Nguyen, Terrence Li, Suyesh Shrestha, Megan Kim Vu, Jerry Yining Wang, and Vincent Ng. 2024. [MemeIntent: Benchmarking intent description generation for memes](#). In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 631–643, Kyoto, Japan. Association for Computational Linguistics.
- Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. [Detecting harmful memes and their targets](#). In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, Findings of ACL, pages 2783–2796. Association for Computational Linguistics.
- Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. [Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4542–4550.
- Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. [What do you meme? generating explanations for visual semantic role labelling in memes](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9763–9771.
- Shivam Sharma, Tharun Suresh, Atharva Kulkarni, Himanshi Mathur, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2022. [Findings of the CONSTRAINT 2022 shared task on detecting the hero, the villain, and the victim in memes](#). In *Proceedings of the Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situations*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics.
- Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. [Drivelm: Driving with graph visual question answering](#). In *European Conference on Computer Vision*, pages 256–274. Springer.
- Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.
- 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*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. [Causal mediation analysis for interpreting neural NLP: the case of gender bias](#). *CoRR*, abs/2004.12265.
- Mo Wang, Kaixuan Ren, Pratik Jalan, Ahmed Ashraf, Tuong Vy Vu, Rahul Seetharaman, Shah Nawaz, and Usman Naseem. 2026a. [From native memes to global moderation: Cross-cultural evaluation of vision-language models for hateful meme detection](#). In *Proceedings of the ACM Web Conference 2026*, WWW ’26, page 9788–9799, New York, NY, USA. Association for Computing Machinery.
- Tongguan Wang, Wei Chen, Feiyue Xue, Junkai Li, Hui Liu, Xiqiao Ba, Yang Xiao, and Ying Sha. 2026b. [Dfm-nc: Disentangling fine-grained multiplex non-literal cues for multimodal sentiment analysis](#). *Expert Systems with Applications*, page 133778.
- Tongguan Wang, Junkai Li, Guixin Su, Yongcheng Zhang, Dongyu Su, Yuxue Hu, and Ying Sha. 2025. [Rclmufn: Relational context learning and multiplex fusion network for multimodal sarcasm detection](#). *Knowledge-Based Systems*, 319:113614.
- Tongguan Wang, Junkai Li, Feiyue Xue, Dongyu Su, Guixin Su, Xiaopeng Wen, and Ying Sha. 2026c. [Big-tm: Bridging individual guidance with trifusion mopoe for chinese memes understanding](#). *Knowledge-Based Systems*, 345:116137.
- John S Wilkins. 1998. What’s in a meme? *Journal of Memetics-Evolutionary Models of Information Transmission*, 2(1).
- Liyuan Zhang, Zeyun Cheng, Zhongyan Gui, Yan Yang, Yong Liu, and Jinke Ma. 2025. [T3svfnd: Towards an evolving fake news detector for emergencies with test-time training on short video platforms](#). In *International Conference on Database Systems for Advanced Applications*, pages 339–354. Springer.

Zhengyi Zhao, Shubo Zhang, Yuxi Zhang, Yanxi Zhao, Yifan Zhang, Zezhong Wang, Huimin Wang, Yutian Zhao, Bin Liang, Yefeng Zheng, Binyang Li, Kam-Fai Wong, and Xian Wu. 2025. [MemeReaCon: Probing contextual meme understanding in large vision-language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3559–3582, Suzhou, China. Association for Computational Linguistics.

Appendix

A Details of Dataset Construction

A.1 Dataset Construction Details

MemeScope is constructed to support fine-grained analysis of LVLM meme understanding, with a particular focus on non-literal pragmatic inference. We begin with candidate memes sampled from four existing multimodal meme datasets: MultiOFF (Suryawanshi et al., 2020), ExHVV (Sharma et al., 2023), MemeCap (Hwang and Schwartz, 2023), and MemeReaCon (Zhao et al., 2025). These sources provide diverse meme styles, topics, and image-text configurations. We briefly describe each source below.

- **MultiOFF** (Suryawanshi et al., 2020): A multimodal offensive meme dataset designed for detecting offensive content from paired image and text. It provides socially grounded meme examples where meaning often depends on the interaction between visual and textual cues.
- **ExHVV** (Sharma et al., 2023): A dataset focusing on visual-verbal meme interpretation. It contains memes that require understanding how image and text jointly convey implicit meanings, making it useful for collecting candidates with non-literal or figurative content.
- **MemeCap** (Hwang and Schwartz, 2023): A captioning-oriented meme understanding dataset. It emphasizes generating explanations or captions for memes, and therefore contains many examples involving visual metaphors, background knowledge, and implicit semantic relations.
- **MemeReaCon** (Zhao et al., 2025): A reasoning-oriented meme dataset that highlights contextual and commonsense reasoning in meme interpretation. It provides examples where understanding the intended meaning requires going beyond surface-level image-text matching.

Source Dataset	Count	Percentage
ExHVV (Sharma et al., 2023)	501	28.1%
MemeCap (Hwang and Schwartz, 2023)	796	44.7%
MemeReaCon (Zhao et al., 2025)	359	20.2%
MultiOFF (Suryawanshi et al., 2020)	125	7.0%
Total	1,781	100.0%

Table 4: Source dataset composition of MemeScope after filtering and human verification.

We apply a filtering process to ensure that each retained meme requires more than literal image recognition or text reading. Specifically, we keep samples that satisfy the following criteria:

Filtering Criteria for MemeScope Construction

- It contains both visual and textual cues.
- Its intended meaning cannot be recovered from either modality alone.
- Its interpretation involves a pragmatic mechanism, such as irony, exaggeration, wordplay, cultural reference, absurdity, or image-text incongruity.
- The image is readable enough for both visual and textual interpretation.

We remove duplicates, near-duplicates, low-resolution images, memes with unreadable text, and samples whose meanings are purely literal.

Table 4 reports the source composition of MemeScope. We construct MemeScope by sampling from four existing multimodal meme datasets: MultiOFF (Suryawanshi et al., 2020), ExHVV (Sharma et al., 2023), MemeCap (Hwang and Schwartz, 2023), and MemeReaCon (Zhao et al., 2025). These sources provide complementary coverage of offensive content, visual-verbal interpretation, caption-style explanation, and reasoning-oriented meme understanding. After filtering and human verification, MemeScope contains 1,781 memes in total. MemeCap contributes the largest portion, while MultiOFF contributes a smaller but useful set of offensive and socially grounded memes.

A.2 Pragmatic Mechanism Definitions

Table 5 provides the definitions of the eight pragmatic mechanism labels used in MemeScope. When multiple mechanisms are present, annotators select the dominant mechanism that best explains the meme’s core communicative effect.

Mechanism	Definition
Irony	The meme conveys a meaning that contrasts with its literal wording or apparent situation. The intended stance differs from the surface expression.
Exaggeration	The meme amplifies a situation, emotion, trait, or consequence beyond its realistic scale, relying on overstatement or hyperbole.
Juxtaposition	The meme places contrasting elements side by side, such as image versus text, expectation versus reality, or past versus present. The meaning arises from the contrast.
Progressive Reveal	The meme discloses information in stages, causing an initial interpretation to be revised or overturned by later text or visual content.
Wordplay	The meme relies on linguistic manipulation, such as puns, phonetic similarity, spelling variation, portmanteau, or double meaning.
Objectification	The meme maps abstract ideas, people, emotions, groups, or situations onto concrete visual entities through labels or symbolic assignment.
Absurdity	The meme presents an impossible, surreal, illogical, or commonsense-violating situation, with the effect arising from its irrationality.
Cultural Reference	The meme depends on external cultural, social, historical, political, or community-specific knowledge that must be recognized to infer the intended meaning.

Table 5: Definitions of the eight pragmatic mechanism labels in MemeScope.

A.3 Annotation Prompts

After filtering, each retained meme is converted into a structured annotation instance. Each instance contains the meme image, extracted or provided meme text, a factual text description, a counterfactual text description, and stage-wise annotations for visual grounding, cross-modal alignment, and pragmatic inference. The counterfactual text is designed to preserve literal content while neutralizing the original pragmatic cue, making it suitable for the hidden-state patching experiments in Section 4.4.

To support reproducibility, we provide the annotation schema and an abbreviated version of the LLM-assisted annotation prompt. The full prompt with complete two-shot examples will be released together with MemeScope.

Prompt for LLM-assisted MemeScope Annotation

You are an expert annotator in multimodal cognitive linguistics and internet culture. Your task is to analyze the provided meme image, extract its original text, and output a strict JSON object following a 3-Stage Cognitive Decomposition protocol.

Protocol Definitions:

1. **text_factual**: The exact original text in the meme.
2. **text_counterfactual**: A literal, boring, and logical description of the visual scene, with metaphors, irony, emotional conflict, and other pragmatic effects removed.
3. **stage1_grounding**: Objective description of physical entities, scenes, actions, and visible text.
4. **stage2_alignment**: Mapping between textual cues and visual elements, including any literal conflict or incongruity.

5. **stage3_pragmatics**: The implicit mechanism, required background knowledge, and pragmatic mechanism.

Candidate pragmatic mechanisms: *Irony, Exaggeration, Juxtaposition, Progressive Reveal, Wordplay, Objectification, Absurdity, and Cultural Reference.*

Final Output Instructions: Extract the meme text accurately using | as a separator. Return only a raw JSON object with the required fields. Do not include markdown formatting.

Required Annotation Schema

```
{
  "meme_id": "...",
  "image_path": "...",
  "text_factual": {
    "text": "..."
  },
  "text_counterfactual": {
    "text": "..."
  },
  "annotations": {
    "stage1_grounding": {
      "literal_visuals": "..."
    },
    "stage2_alignment": {
      "text_image_mapping": "...",
      "literal_conflict": "..."
    },
    "stage3_pragmatics": {
      "implicit_mechanism": "...",
      "background_knowledge": "...",
      "pragmatic_intent": "..."
    }
  }
}
```

Abbreviated Annotation Example

```
{
  "meme_id": "memescope_0002.png",
  "image_path": "memescope_images/memescope_0002.png",
  "text_factual": {
    "text": "Does writing code make you happy? | YES | YESTERDAY IT ONLY MADE ME CRY 3 TIMES"
  },
  "text_counterfactual": {
    "text": "A sword is being drawn | YES | The blade contains a longer written message"
  }
}
```

Does writing
code make you happy?

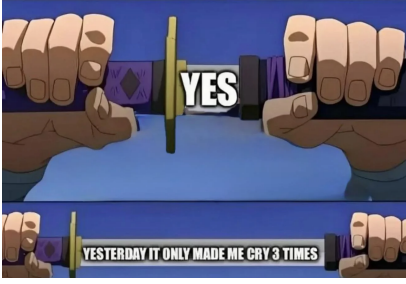


Figure 7: Example meme from MemeScope (memescope_0002.png) with corresponding stage-wise annotations.

```

},
"annotations": {
  "stage1_grounding": {
    "literal_visuals": "The image shows a sword-drawing  
meme template. A question appears above the image,  
and additional text is gradually revealed on the  
sword blade."
  },
  "stage2_alignment": {
    "text_image_mapping": "'YES': 'initial visible  
answer on the blade', 'YESTERDAY IT ONLY MADE ME CRY  
3 TIMES': 'hidden continuation revealed later'",
    "literal_conflict": "A sword cannot literally answer  
a question about coding happiness. The humor comes  
from the contrast between the short positive answer  
and the painful hidden qualification."
  },
  "stage3_pragmatics": {
    "implicit_mechanism": "Progressive Reveal",
    "background_knowledge": "Programming is often  
described online as rewarding but emotionally  
frustrating.",
    "pragmatic_intent": "The meme humorously presents  
coding as enjoyable while revealing that it still  
causes repeated frustration."
  }
}
}

```

Figure 7 shows a representative meme from MemeScope, along with its detailed stage-wise annotations. This example illustrates how visual grounding, cross-modal alignment, and pragmatic inference are annotated:

- **Visual Grounding (Stage 1):** Literal entities and actions are described, including the sword-drawing meme template and the visual elements gradually revealed on the blade.
- **Cross-modal Alignment (Stage 2):** Text-image mapping and conflicts are annotated, highlighting where textual cues contrast with visual content to produce non-literal meaning.
- **Pragmatic Inference (Stage 3):** Implicit mechanism, background knowledge, and prag-

matic mechanism are annotated, capturing the non-literal humor and underlying narrative.

This example demonstrates the granularity and depth of annotations in MemeScope, allowing us to analyze where LVLMs correctly ground literal content but may still fail at pragmatic inference.

A.4 Annotation Quality Control

MemeScope follows an LLM-assisted, human-verified annotation pipeline. The LLM is used to generate initial structured drafts, but all final annotations are determined by human reviewers. Three Ph.D. candidates review every meme instance. During verification, reviewers check the correctness of visual descriptions, text-image mappings, pragmatic mechanism labels, background knowledge, pragmatic mechanisms, and counterfactual text.

The review process focuses on three common sources of error. First, reviewers correct visual hallucinations, such as objects or actions that are not present in the image. Second, they revise incorrect or incomplete text-image mappings, especially when the meme relies on a subtle visual-textual relation. Third, they adjudicate pragmatic mechanism labels when multiple mechanisms appear plausible. In such cases, reviewers select the dominant mechanism that best explains the meme’s intended communicative effect.

Disagreements are resolved through discussion. If no single mechanism can be reliably determined, the sample is removed from the dataset. This conservative strategy helps ensure that MemeScope supports stable evaluation of pragmatic mechanism understanding. To quantify annotation reliability, we compute inter-annotator agreement on pragmatic mechanism labels using Fleiss’ Kappa. The agreement score was $\kappa = 0.72$, indicating substantial agreement among annotators.

B Experimental Setup

B.1 Models

We evaluate three representative open-weight large vision-language models (LVLMs) in this work. These models cover different LVLm families and allow us to test whether the observed interpretability patterns persist across architectures. All LVLm parameters are kept frozen throughout evaluation, probing, and patching. Their key characteristics are summarized below.

- **LLaVA-1.5-7B** (Liu et al., 2023): Based on the Vicuna language backbone, this model has

32 transformer layers and approximately 7B parameters. It integrates visual input through a visual encoder and multimodal adapters to enable joint image-text understanding.

- **Qwen2.5-VL-7B** (Bai et al., 2025b): A 7B parameter LVLM with 28 transformer layers. It uses a pre-trained Qwen language model as a backbone and incorporates multimodal fusion modules for visual grounding and cross-modal alignment. It supports high-quality multimodal reasoning across images and text.
- **Qwen3-VL-8B** (Bai et al., 2025a): An 8B parameter LVLM with 36 transformer layers. Extends Qwen2.5-VL with larger model capacity and deeper layers for enhanced multimodal integration. Designed for robust reasoning on complex image-text tasks, including pragmatic inference in memes.

B.2 Evaluation Protocol

We use stratified 5-fold cross-validation on MemeScope. In each fold, 80% of the samples are used for training or development, and the remaining 20% are held out for testing. The folds preserve the distribution of the eight pragmatic mechanisms. Direct LVLM evaluation is performed without fine-tuning; the development split is used only for prompt calibration. Layer-wise probes are trained on the training split and evaluated on the corresponding held-out split. Hidden-state patching is conducted on held-out examples.

B.3 Metrics

We use **Macro-F1** as the primary metric for the final eight-way pragmatic mechanism classification task, since MemeScope is label-imbalanced. We also report Accuracy for completeness.

The stage-wise scores in Table 3 are diagnostic metrics for intermediate comprehension abilities and are not identical to the final Macro-F1. Specifically, **Ground. F1** measures performance on the visual grounding diagnostic task, and **Align. F1** measures performance on the cross-modal alignment diagnostic task. Both are reported as Macro-F1 over their corresponding stage-specific label spaces.

For layer-wise probing, we report Macro-F1 at each layer to measure how linearly decodable each stage-specific signal is from hidden representations. For hidden-state patching, we report the total effect, indirect effect, and normalized causal effect ratio

defined in Section 4.4, which quantify whether internal representations causally affect the final pragmatic prediction.

B.4 Implementation Details

All models are evaluated with a unified multiple-choice prompt over the same eight pragmatic mechanisms. We consider full multimodal input, OCR/Text-only input, Image-only input, image with counterfactual text, and shuffled image-text pairs for modality analysis. For probing, we extract hidden states from every transformer layer and train independent linear probes without updating the LVLM. For causal patching, each factual input x_f is paired with a counterfactual input x_c , and we replace the counterfactual hidden state with the corresponding factual hidden state following $h_c^{(l)} \leftarrow h_f^{(l)}$. All experiments are conducted on a single Nvidia A100-PCIE-40G GPU using Python 3.10.18 and PyTorch 2.5.1.

C Experimental Details

C.1 Candidate Label Set

Each meme is evaluated as an eight-way pragmatic mechanism classification problem. The candidate label set is:

$$\mathcal{Y} = \{Irony, Exaggeration, Juxtaposition, Progressive Reveal, Wordplay, Objectification, Absurdity, Cultural Reference\}. \quad (11)$$

Models are prompted to select exactly one label that best describes the core pragmatic mechanism of the meme.

C.2 Prompting and Output Normalization

We use the same multiple-choice prompt across all LVLMs. The prompt presents the meme image, the meme text when available, and the eight candidate mechanisms. Model outputs are first parsed by exact label matching. If the output contains a synonymous term, we apply deterministic alias normalization, such as mapping “pun” to *Wordplay*. Outputs that cannot be mapped to any candidate label are treated as invalid predictions.

Test Prompt for LVLM

```
USER: <image>
Text on meme: {meme_text}
Which of the following 8 pragmatic mechanisms best
describes the core mechanism of this meme?
A) Irony/Sarcasm
```

B) Exaggeration/Hyperbole
 C) Juxtaposition/Contrast
 D) Progressive Reveal
 E) Wordplay/Pun
 F) Objectification/Labeling
 G) Absurdity/Surrealism
 H) Cultural/Stereotype Reference
 Answer with the correct option letter directly.
 ASSISTANT: The best option is

The same candidate label set is used for all models. During evaluation, outputs are normalized by exact label matching and a small set of deterministic label aliases, such as mapping “pun” to *Wordplay* when the model uses a synonymous term.

C.3 Input Settings

For modality analysis, we evaluate five input settings. *Full input* provides the image, OCR/text, and task prompt. *OCR/Text-only* removes the image and keeps only the meme text and prompt. *Image-only* provides the image and prompt without separately supplying OCR/text. *Image + counterfactual text* replaces the original meme text with a literalized counterfactual version. *Shuffled pair* randomly pairs an image with text from another meme.

C.4 Layer-wise Probing

For each LVLM, we extract hidden states from every transformer layer while keeping the model frozen. Unless otherwise stated, the hidden state at the final input position is used as the default representation. Each layer is equipped with an independent linear probe trained with cross-entropy loss on the training split of each fold. Probe performance is reported on the held-out split and averaged across folds.

C.5 Hidden-state Patching

For causal localization, each example is paired with a factual input x_f and a counterfactual input x_c . The factual input preserves the original pragmatic cue, while the counterfactual input removes or neutralizes it while retaining literal content as much as possible. During patching, we run the model on x_c and replace the target hidden state with the corresponding factual hidden state from x_f :

$$h_c^{(l)} \leftarrow h_f^{(l)}.$$

We apply this intervention across layers to obtain a layer-wise causal localization curve. We also patch different token groups, including visual tokens, OCR/Text tokens, prompt tokens, and fused multimodal states.

C.6 Diagnostic Label Spaces

The stage-wise scores reported in Table 3 are diagnostic classification metrics rather than free-form explanation-generation metrics. Although MemeScope provides natural-language annotations for visual grounding, cross-modal alignment, and pragmatic inference, we derive discrete diagnostic targets for visual grounding and cross-modal alignment to support controlled evaluation and layer-wise analysis. These diagnostic targets are used to measure whether models capture the intermediate evidence and cross-modal relations needed for meme interpretation.

For each meme instance, the free-form annotations for visual grounding and cross-modal alignment are mapped to one categorical label in the corresponding diagnostic label space. The mapping is performed during the human verification stage: annotators first inspect the structured annotation generated by the LVLM-assisted pipeline, and then assign or correct the dominant diagnostic label for each stage. When more than one label is plausible, annotators select the label that best captures the primary evidence needed to interpret the meme. Instances for which no stable dominant label can be determined are removed from the dataset.

Visual Grounding. The visual grounding target captures the dominant type of visual evidence that must be recognized before higher-level meme interpretation is possible. We use the following six labels: *Human/Character-centric*, *Object-centric*, *Scene/Event-centric*, *Text-layout/Template-centric*, *Symbolic/Metaphorical visual cue*, and *Mixed visual evidence*. This target does not evaluate whether the model generates the exact natural-language visual description; instead, it evaluates whether the model representation encodes the correct type of visual evidence.

Cross-modal Alignment. The cross-modal alignment target captures the dominant relation between the meme text and the visual content. We use the following six labels: *Text-image reinforcement*, *Text-image contrast/incongruity*, *Label-to-entity mapping*, *Sequential/reveal-based alignment*, *Visual instantiation of abstract text*, and *External-reference alignment*. This target is intended to diagnose whether the model captures how textual cues interact with visual elements, rather than whether it produces a full natural-language explanation.

Diagnostic task	Prediction target	Label space size
Visual Grounding	Dominant visual evidence type	6
Cross-modal Alignment	Dominant text-image relation	6
Final Pragmatic Classification	Dominant pragmatic mechanism	8

Table 6: Diagnostic and final-task label spaces used in MemeScope evaluation. Ground. F1 and Align. F1 in Table 3 are diagnostic Macro-F1 scores over the first two label spaces. The final pragmatic mechanism label space is evaluated by Acc. and Macro-F1 in Table 3.

Final Pragmatic Mechanism. The final pragmatic mechanism target is the eight-way label used for MemeScope’s main classification task: *Irony*, *Exaggeration*, *Juxtaposition*, *Progressive Reveal*, *Wordplay*, *Objectification*, *Absurdity*, and *Cultural Reference*. This target measures whether the model identifies the dominant non-literal mechanism underlying the meme’s communicative effect.

D Per-class Performance Analysis

To further analyze the difficulty of different pragmatic mechanisms, we report per-class F1 scores under the Full input setting in Table 7. The average of the eight per-class F1 scores corresponds to the Macro-F1 reported in Table 3. This analysis complements the overall Macro-F1 results by showing whether model performance is uniformly distributed across pragmatic mechanisms or concentrated on specific categories.

As shown in Table 7, the overall trend is consistent with the main results in Table 3: Qwen3-VL-8B achieves the strongest performance across most pragmatic mechanisms, followed by Qwen2.5-VL-7B and LLaVA-1.5-7B. Among the eight categories, Progressive Reveal and Absurdity are consistently more challenging, suggesting that current LVLMs struggle with staged interpretation shifts and commonsense-violating situations that require reasoning beyond literal image-text cues. In contrast, Cultural Reference and Irony obtain relatively higher scores, indicating that LVLMs can more effectively exploit salient textual or contextual priors when the pragmatic mechanism is explicitly signaled. These results show that the difficulty of MemeScope is not uniform across categories, but is especially pronounced for mechanisms requiring fine-grained temporal, counterfactual, or non-literal reasoning.

E Additional Token-group Analyses

Figures 8 and 9 present the token-group causal effect heatmaps for Qwen2.5-VL-7B and Qwen3-VL-8B, respectively. The patterns are consistent with

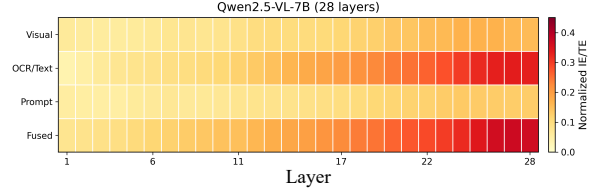


Figure 8: Layer $\times$ token-group causal effect heatmap for Qwen2.5-VL-7B.

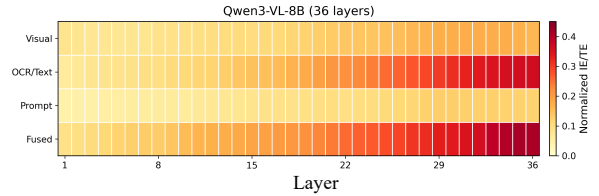


Figure 9: Layer $\times$ token-group causal effect heatmap for Qwen3-VL-8B.

LLaVA-1.5-7B: fused multimodal states and OCR/Text tokens exhibit the strongest causal influence in middle-to-late and late layers, while visual and prompt tokens contribute moderately or minimally. These results confirm that meme-level pragmatic inference primarily depends on late-stage textual and multimodal integration, with minimal reliance on early visual or prompt signals.

F Additional Experimental Results

Table 8 summarizes the peak layer and stage-wise Macro-F1 for visual grounding, cross-modal alignment, and pragmatic inference across the three LVLMs. Consistent patterns emerge: visual grounding is most accessible in early layers, cross-modal alignment peaks in middle layers, and pragmatic inference emerges mainly in late layers. Qwen3-VL-8B shows the strongest representation quality across all stages, followed by Qwen2.5-VL-7B and LLaVA-1.5-7B. Pragmatic representations remain weaker than grounding and alignment, reflecting the inherent difficulty of non-literal pragmatic mechanism prediction.

Table 10 reports the average and peak causal contributions (IE/TE) of different token groups. Across

Model	Irony	Exag.	Juxta.	Prog. Rev.	Wordplay	Object.	Absurd.	Cult. Ref.
LLaVA-1.5-7B	48.6%	43.2%	41.5%	24.8%	38.7%	44.1%	31.6%	61.9%
Qwen2.5-VL-7B	55.4%	51.0%	49.2%	32.7%	45.8%	50.4%	39.9%	65.2%
Qwen3-VL-8B	60.7%	56.1%	54.0%	37.9%	50.6%	55.3%	44.7%	67.9%

Table 7: Per-class F1 scores for eight-way pragmatic mechanism classification on MemeScope under the Full input setting. The average of the eight per-class F1 scores corresponds to the Macro-F1 reported in Table 3.

Model	Ground.	Align.	Prag.
LLaVA-1.5-7B	L7 / 0.51	L16 / 0.41	L31 / 0.28
Qwen2.5-VL-7B	L8 / 0.55	L17 / 0.45	L27 / 0.31
Qwen3-VL-8B	L8 / 0.58	L18 / 0.49	L35 / 0.34

Table 8: Peak layer and stage-wise Macro-F1 for visual grounding (Ground.), cross-modal alignment (Align.), and pragmatic inference (Prag.).

models, fused multimodal states and OCR/Text tokens consistently mediate the strongest effects, especially in late layers. Visual and prompt tokens contribute moderately or minimally. This confirms that meme-level pragmatic inference relies primarily on late-stage textual and multimodal integration rather than early visual cues.

Table 9 presents the modality ablation results measured by Macro-F1. The Full setting achieves the best performance for all three LVLMs, with 41.8%, 48.7%, and 53.4% Macro-F1 for LLaVA-1.5-7B, Qwen2.5-VL-7B, and Qwen3-VL-8B, respectively. OCR/Text-only input consistently outperforms Image-only input across models, suggesting that textual cues provide stronger pragmatic signals than visual cues alone. The Image + CF Text setting remains substantially below the Full setting, indicating that neutralizing the original pragmatic trigger weakens multimodal pragmatic inference. Compared with Image-only input, Image + CF Text yields similar or slightly lower performance for LLaVA-1.5-7B and Qwen2.5-VL-7B, but slightly higher performance for Qwen3-VL-8B, suggesting that counterfactual text may still provide limited scene-level or structural cues. Shuffled image-text pairs also perform below the Full setting but remain above Image-only input for all three models, indicating that LVLMs partially exploit textual priors even when image-text alignment is disrupted.

G Additional Qualitative Cases

We provide additional qualitative cases to examine whether the pragmatic drift pattern observed in Figure 1 also appears across different LVLMs.

Each case reports the ground-truth pragmatic mechanism, the model’s dominant prediction, and the layer-wise shift of probability mass among candidate mechanisms. Rather than treating errors as simple recognition failures, we focus on whether the model first captures surface visual or textual cues but later moves toward a plausible yet incorrect pragmatic interpretation.

G.1 LLaVA-1.5-7B Cases

Figures 10–12 show three qualitative cases from LLaVA-1.5-7B. Consistent with the caterpillar example in the main text, the model does not simply fail to recognize the visual scene or meme text. Instead, it often captures salient surface cues, but its deeper-layer predictions drift toward plausible yet incorrect pragmatic mechanisms.

Figure 10 illustrates a Wordplay failure. The meme relies on the ambiguity of “hot dog”: the phrase can refer to both an actual dog exposed to heat and the food item shown in the search result. LLaVA-1.5-7B initially assigns relatively diffuse probabilities across mechanisms, but the Wordplay signal remains weak throughout the layer stack. In later layers, probability mass moves sharply toward Juxtaposition, which is plausible because the image does contain a contrast between the dog and the food item. However, this interpretation misses the core linguistic trigger. The failure, therefore, reflects pragmatic drift from a pun-based reading to a broader contrast-based explanation.

Figure 11 shows a Juxtaposition case. The humor comes from placing the commercial phrase “wipe and wax” in the context of toilet paper, creating an incongruous cross-modal contrast between product language and bathroom imagery. The model appears to recognize the relevant text and object, but its prediction does not preserve the intended contrast relation. Instead, later layers assign high probability to Exaggeration, with additional instability around Progressive Reveal. This suggests that the model treats the meme as an intensified or exaggerated claim rather than as a clash

Model	Full	OCR/Text	Image	Image + CF Text	Shuffled Pair
LLaVA-1.5-7B	41.8%	34.2%	29.6%	29.6%	31.8%
Qwen2.5-VL-7B	48.7%	39.7%	35.1%	34.7%	36.9%
Qwen3-VL-8B	53.4%	44.1%	38.9%	39.3%	41.1%

Table 9: Modality ablation results on MemeScope measured by Macro-F1. The Full, OCR/Text, and Image settings are consistent with the Macro-F1 scores reported in Table 3. Image + CF Text denotes replacing the original meme text with counterfactual text that neutralizes the core pragmatic trigger. Shuffled Pair denotes randomly pairing meme images and texts from different samples.

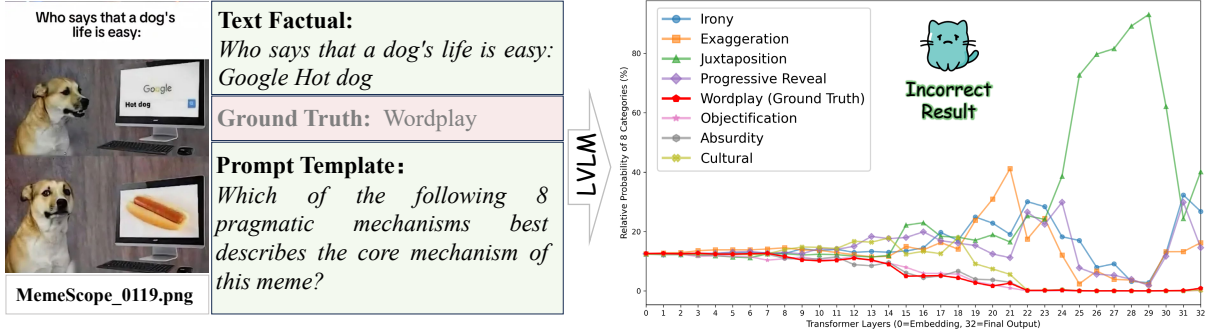


Figure 10: LLaVA-1.5-7B example on a Wordplay meme. The meme relies on the double meaning of “hot dog”, contrasting the literal dog image with the food-related search result. Although Wordplay is the ground-truth mechanism, the model’s later layers shift strongly toward Juxtaposition, treating the meme mainly as a visual-textual contrast rather than a linguistic pun.

Token Group	Avg. IE/TE	Peak Bin	Peak IE/TE
Visual	0.12	l_{late}	0.22
OCR/Text	0.22	l_{late}	0.42
Prompt	0.10	l_{late}	0.21
Fused	0.26	l_{late}	0.48

Table 10: Causal contribution of each token group to pragmatic inference measured by normalized IE/TE. Late-stage OCR/Text and fused states mediate the strongest effects.

between two semantic frames. The error is therefore not a simple grounding failure, but a failure to maintain the correct pragmatic relation between the visual object and the overlaid text.

Figure 12 presents a Progressive Reveal example. The meme builds a staged interpretation through the sequence “git commit”, “git push”, and “git add”, with the final panel humorously violating the expected order of Git operations. LLaVA-1.5-7B shows some sensitivity to the correct mechanism in middle layers, where the Progressive Reveal probability rises. However, this signal is not stable. In the later layers, the model increasingly favors Juxtaposition, likely because the three panels create a strong visual contrast across airplane states. This again matches the pragmatic drift pattern: the model recognizes salient visual progression but

collapses the staged reversal into a more generic contrast-based mechanism.

Overall, these LLaVA-1.5-7B cases show that pragmatic drift occurs across mechanisms. The model often grounds visible entities and reads the text correctly, but this partial understanding is not preserved in later layers. Instead, later representations overemphasize a salient alternative, such as Juxtaposition or Exaggeration, and suppress the intended label. This supports our finding that meme-level pragmatic inference is not only about visual recognition or OCR accuracy. It also requires maintaining the relation between image, text, and background knowledge that defines the intended non-literal mechanism.

G.2 Qwen2.5-VL-7B Cases

Figures 13–15 show additional qualitative cases from Qwen2.5-VL-7B. Compared with LLaVA-1.5-7B, Qwen2.5-VL-7B often assigns higher confidence to a single late-layer mechanism, but this confidence does not always correspond to the correct pragmatic interpretation. These cases suggest that stronger late-layer certainty can still amplify an incorrect reading when the model overweights a salient cultural, absurd, or contrastive cue.

Figure 13 shows an Objectification failure. The meme uses the phrase “Republican Voting Booth”

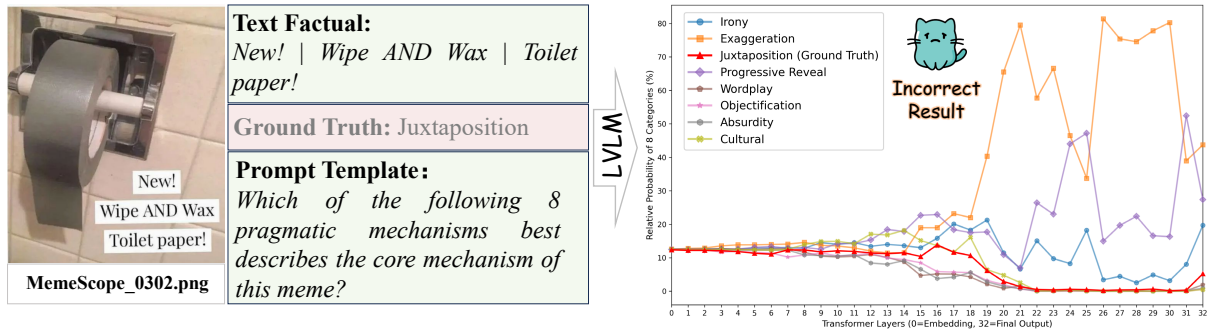


Figure 11: LLaVA-1.5-7B example on a Juxtaposition meme. The ground-truth mechanism is the contrast between the advertised phrase “wipe and wax” and the literal toilet-paper holder. The model instead assigns high late-layer probability to Exaggeration, indicating a drift from cross-modal contrast to an overstatement-like interpretation.

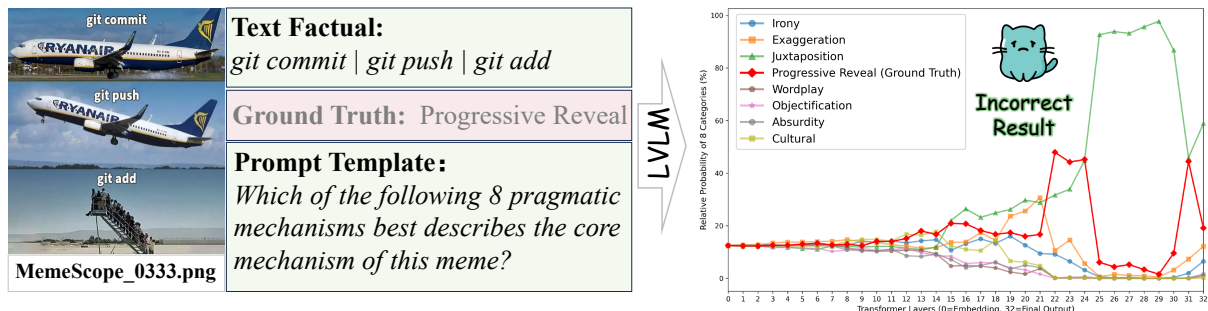


Figure 12: LLaVA-1.5-7B example on a Progressive Reveal meme. The meme maps “git commit”, “git push”, and “git add” onto increasingly absurd airplane scenes, where the final panel reverses the expected command order. The model partially activates the correct mechanism in middle layers, but later drifts toward Juxtaposition.

to map an abstract political group and voting behavior onto a concrete visual setup. The intended mechanism is therefore not simply that the image is strange or politically loaded, but that the visual objects stand in for a social-political concept. Qwen2.5-VL-7B partially captures the political context, as reflected by the late increase of Cultural Reference. However, the correct Objectification signal remains weak, and the model assigns substantial probability to Absurdity and Cultural Reference in later layers. This indicates a drift from symbolic mapping to a more generic reading of political context and visual strangeness.

Figure 14 presents an Absurdity case. The meme becomes humorous because it places the invented or context-dependent term “covfefe” alongside substances such as cocaine, marijuana, and beer, producing an irrational comparison that violates ordinary semantic expectations. Qwen2.5-VL-7B recognizes this absurd relation more clearly than in the previous case: the Absurdity curve rises strongly in the later layers and becomes dominant for several steps. Nevertheless, the model remains unstable near the end, where competing mechanisms such

as Juxtaposition also receive high probability. This case is less a complete failure of recognizing absurdity than a failure to preserve a stable final decision under competing surface patterns.

Figure 15 illustrates a Juxtaposition failure. The meme contrasts a broad social label, “Society”, with a specific behavioral description about carrying trash until finding a trash can. The intended effect comes from placing a mundane moral behavior against a broad social frame. Qwen2.5-VL-7B maintains some probability for Juxtaposition in earlier and middle layers, suggesting that it detects the contrastive setup. However, in later layers the model increasingly favors Absurdity and then Cultural Reference, likely because the word “Society” encourages a broader social or cultural reading. The final interpretation, therefore, drifts away from the local contrast that defines the meme.

Overall, these Qwen2.5-VL-7B cases show that pragmatic drift is not merely a weakness of older LLMs. Even when the model captures important cues, its late-layer representations can over-commit to a salient but incorrect mechanism. This supports the paper’s claim that meme understanding failures

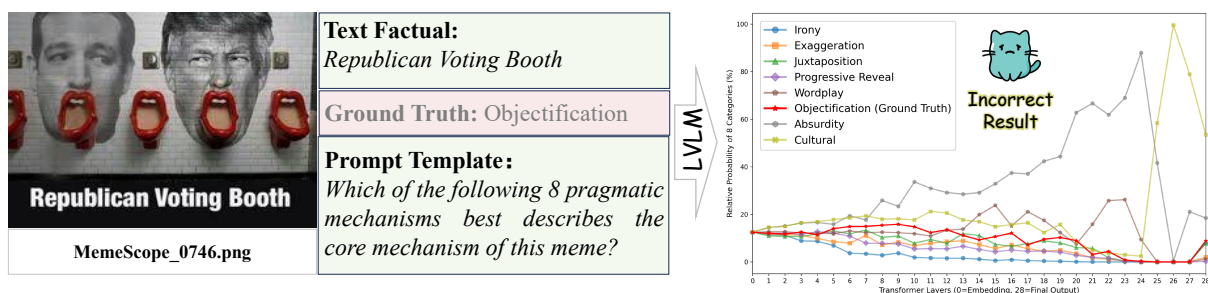


Figure 13: Qwen2.5-VL-7B example on an Objectification meme. The meme labels a voting booth as “Republican Voting Booth” and visually maps political participation onto a concrete object. Although Objectification is the ground-truth mechanism, the model’s late layers drift toward Absurdity and Cultural Reference, treating the meme mainly as a surreal or politically contextualized image rather than a symbolic mapping.

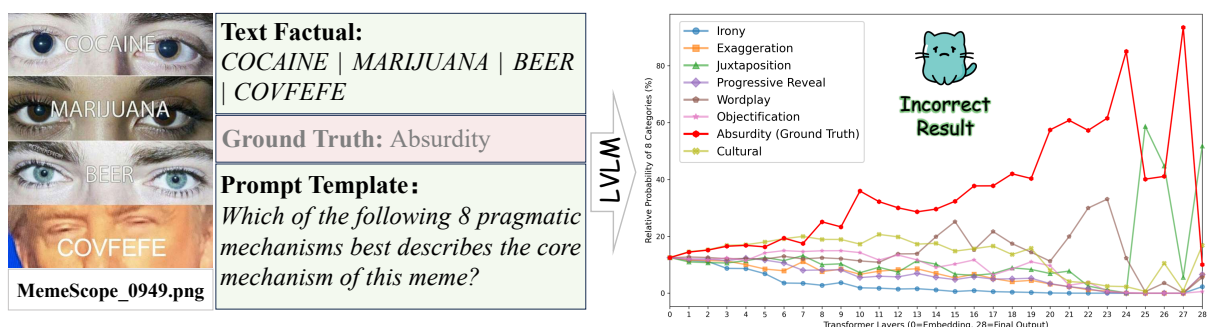


Figure 14: Qwen2.5-VL-7B example on an Absurdity meme. The meme lists substances such as “cocaine”, “marijuana”, “beer”, and “covfe” over eye images, creating an illogical and surreal comparison. The model activates the correct Absurdity mechanism strongly in later layers, but the final prediction remains unstable and is affected by competing mechanisms such as Juxtaposition.

often arise after partial grounding, when the model must select the intended non-literal relation among plausible alternatives.

G.3 Qwen3-VL-8B Cases

Figures 16–18 show qualitative cases from Qwen3-VL-8B. Although this model generally produces stronger layer-wise signals than LLaVA-1.5-7B and Qwen2.5-VL-7B, the examples show that stronger confidence does not necessarily imply more stable pragmatic inference. In several cases, the model captures salient textual or visual cues, but deeper layers amplify an incorrect mechanism and suppress the ground-truth label.

Figure 16 illustrates an Absurdity failure. The meme relies on a self-contradictory emotional state: the speaker is afraid of both living and dying, which creates an impossible and irrational situation. Qwen3-VL-8B initially assigns non-negligible probability to several mechanisms, including Absurdity, but the correct signal remains weak and is eventually suppressed. In later layers, the model shifts strongly toward Juxtaposition, likely because it treats the two opposing statements

as a contrastive pair. This is a plausible surface reading, but it misses the core absurdity of the self-defeating situation. The case, therefore, reflects a drift from logical impossibility to generic contrast.

Figure 17 shows an Irony case. The humor depends on the well-known mismatch between the names “Greenland” and “Iceland” and their expected visual meanings. The meme uses this reversal to create an ironic setup, strengthened by the “invading army” label. Qwen3-VL-8B captures some of the relevant background contrast in the middle layers, where Irony receives noticeable probability. However, the model later assigns almost all probability to Absurdity. This suggests that it recognizes the scene as strange, but fails to preserve the ironic relation between naming expectations and visual reality. The error matches the pragmatic drift pattern: a culturally grounded irony is reduced to a more generic absurd scenario.

Figure 18 presents a Progressive Reveal failure. The meme unfolds through a sequence: the teacher assigns work, promises class time, does not provide enough time, and then asks why no one finished.

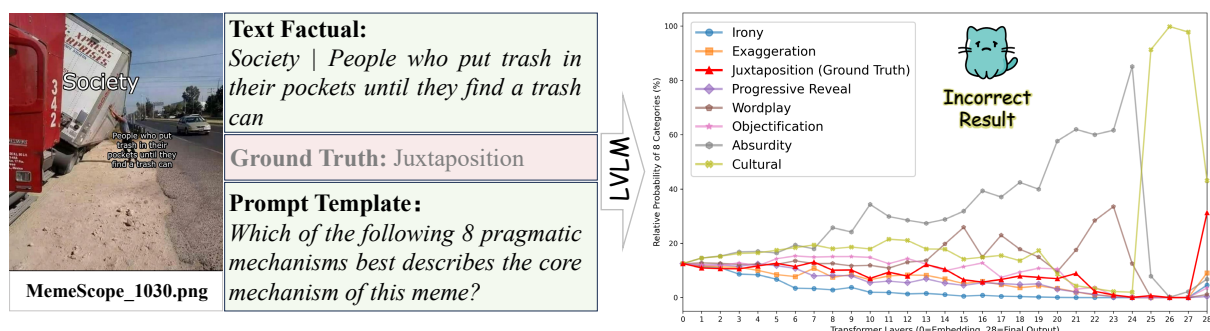


Figure 15: Qwen2.5-VL-7B example on a Juxtaposition meme. The meme contrasts “Society” with people who keep trash in their pockets until they find a trash can. The model recognizes some contrastive structure, but late-layer probability shifts strongly toward Cultural Reference and Absurdity, leading to an incorrect result.

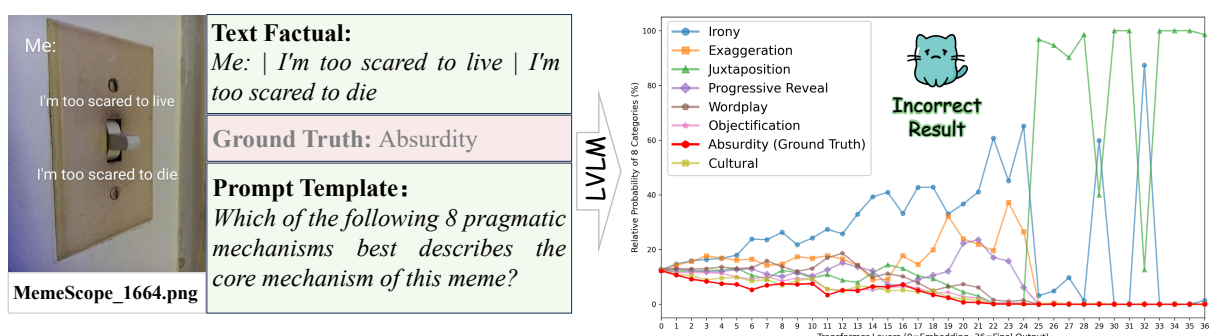


Figure 16: Qwen3-VL-8B example on an Absurdity meme. The meme presents an existentially contradictory statement, “I’m too scared to live” and “I’m too scared to die”, over a light switch, creating an illogical and self-defeating situation. Although Absurdity is the ground-truth mechanism, the model’s late layers drift toward Juxtaposition and Irony.

The intended mechanism is Progressive Reveal because each step revises the interpretation of the previous one and exposes the contradiction in the teacher’s behavior. Qwen3-VL-8B assigns some probability to the correct label in the middle layers, showing partial sensitivity to the staged structure. However, this signal is not sustained. In deeper layers, the model strongly favors Exaggeration and later Juxtaposition, likely because the clown transformation and teacher-student contrast are visually salient. Thus, the model captures the surface progression but loses the specific reveal-based relation that defines the meme.

Overall, these Qwen3-VL-8B examples show that pragmatic drift persists in a stronger LVLM. The model often detects meaningful cues, such as contradiction, cultural reversal, or staged narrative structure, but its late-layer representations can still over-select a salient neighboring mechanism. This supports the paper’s conclusion: meme understanding failures often arise not from missing all evidence, but from failing to maintain the intended pragmatic relation through deeper inference.

These examples show that pragmatic drift is not limited to one model or mechanism: LVLMs can partially ground meme content yet fail to preserve the intended non-literal relation. This provides qualitative evidence for Sections 6.2 and 6.3, where pragmatic information emerges late and is mediated mainly by higher-layer fused and textual representations.

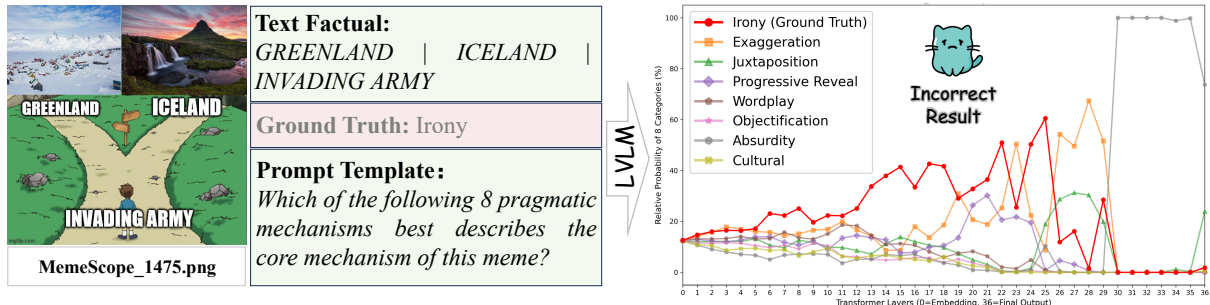


Figure 17: Qwen3-VL-8B example on an Irony meme. The meme contrasts the names “Greenland” and “Iceland” with their opposite visual associations, then adds an “invading army” label. The ground-truth mechanism is Irony, but the model’s late layers shift toward Absurdity, treating the meme as an implausible scenario rather than an ironic reversal.

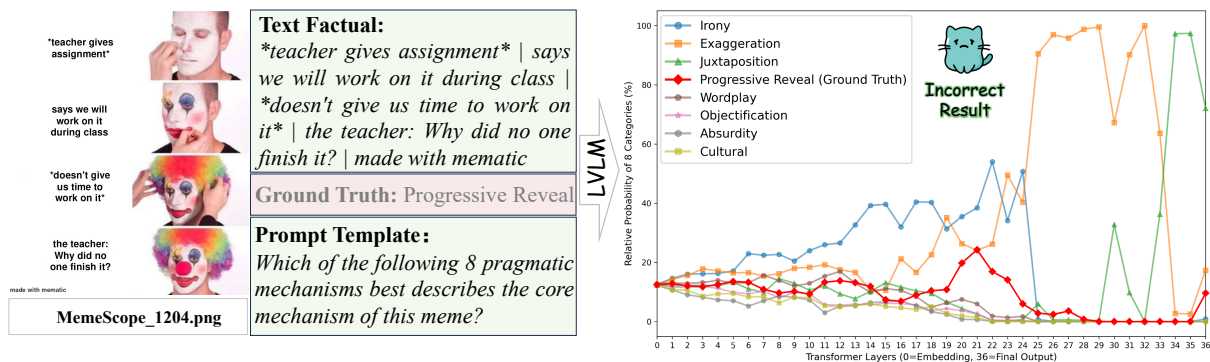


Figure 18: Qwen3-VL-8B example on a Progressive Reveal meme. The meme gradually reveals the contradiction between the teacher’s promise of in-class work time and the later complaint that no one finished. The model detects parts of the staged setup, but late-layer predictions drift toward Exaggeration and Juxtaposition.