SARCATLAS: FINE-GRAINED ANNOTATION AND EFFICIENT MODELING OF MULTIMODAL SARCASM

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sarcasm detection is commonly framed as binary classification. However, a correct prediction does not establish that the model understands why a sample is sarcastic. Meanwhile, annotation policy shifts across source datasets can induce dataset-specific decision criteria and limit cross-dataset generalization. We introduce SarcAtlas, a unified set of fine-grained annotations for MMSD, MMSD2.0, and the English and Chinese splits of SarcNet. SarcAtlas preserves each benchmark's original instances, splits, and labels. It adds annotations for strict and broad sarcasm, type and target, shared and modality-specific information, cross-modal relations, rationales, and interpretation dependencies. We control annotation quality through staged annotation, review, repair, and expert audit. Our analysis reveals systematic differences among benchmarks in label boundaries, pragmatic types, evidence structures, and interpretation requirements. Under the Original labels, cross-dataset evaluation mixes content shifts with annotation policy shifts. Based on these annotations, we develop a single-tower Qwen3-VL-Embedding model with five latent-token groups assigned to different annotated fields and a decision token. Training combines multi-task prediction, reconstruction of the annotated text fields, distillation of teacher-defined sample similarities, and label-supervised contrastive learning. At inference, a single image–text prefill supports binary detection and fine-grained prediction through parallel heads without autoregressive generation. The latent tokens are decoded only for case inspection. We evaluate in-domain detection, cross-dataset and cross-lingual transfer, case-level interpretability, and inference characteristics. Overall, SarcAtlas provides common decision boundaries for comparing multimodal sarcasm models, while its latent tokens make intermediate semantic information available for case-level inspection without sacrificing classification-based inference efficiency.

1 INTRODUCTION

Multimodal sarcasm hinges on a gap between surface content and intended attitude. This gap may occur within the text or emerge jointly from the image, the text, and relevant background knowledge. Its boundary is further blurred by neighboring phenomena such as mockery, direct criticism, and ordinary humor. Multimodal sarcasm detection should therefore answer not only *whether* a post is sarcastic, but also *why* the judgment holds.

Prior work has advanced this task through both dataset construction and cross-modal modeling. MMSD established a large-scale benchmark and a hierarchical fusion framework (Cai et al., 2019). MMSD2.0 subsequently reduced spurious dataset cues and strengthened multi-view image–text modeling (Qin et al., 2023), while SarcNet extended the task to English and Chinese settings (Yue et al., 2024). Later approaches have continued to improve attention, contrastive learning, and image–text incongruity modeling, yet their primary supervision remains a single binary label. Detailed related comparisons are provided in Appendix A.

A binary label compresses variation in sarcastic mechanisms, pragmatic boundaries, and supporting evidence. As Figure 1 illustrates, the same original positive label spans strict sarcasm, broad-only ridicule, direct criticism, and content that merely mentions sarcasm. Under such supervision, a prediction may reflect genuine understanding of implicit attitude, but it may also result from word frequencies, visual templates, or dataset bias. Likewise, a cross-dataset performance drop under

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

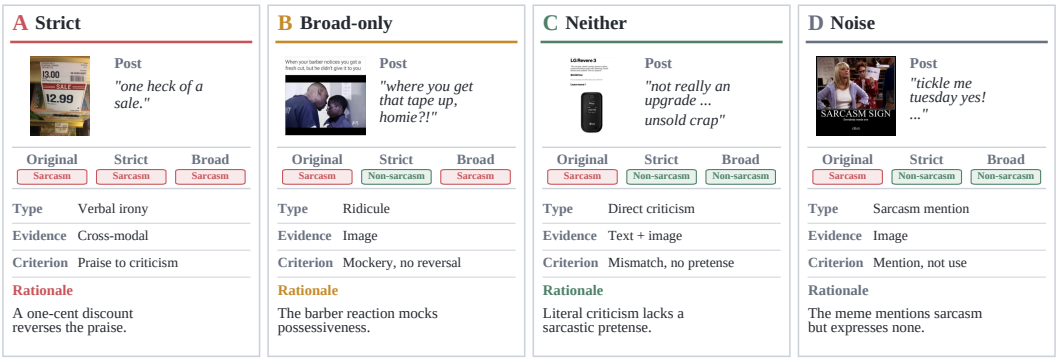


Figure 1: One original positive label spans strict sarcasm, broad-only ridicule, direct criticism, and noise that merely mentions sarcasm. SARCASM/NON-SARCASM indicates the decision under each labeling view; *Original* denotes the original dataset label.

Original labels cannot be cleanly attributed to language, content, or an annotation policy shift. Model representations are becoming richer, while the supervision supplied by the task remains limited.

We address this mismatch by constructing SarcAtlas, a unified set of fine-grained annotations for MMSD, MMSD2.0, and the English and Chinese portions of SarcNet. SarcAtlas preserves the original instances, data splits, and *Original* labels while adding *Strict*/*Broad* boundaries, pragmatic types and targets, shared and modality-unique content, cross-modal relations, sarcasm explanations, and dependencies on OCR, hashtags, cultural context, and external knowledge. Staged annotation, evidence review, targeted repair, and expert audit jointly control annotation quality. Each record contains the final judgment, the supporting image–text evidence, and an explanation of why the sample does or does not qualify as sarcasm.

Building on these annotations, we propose a single-tower model based on Qwen3-VL-Embedding. Five groups of latent tokens are appended to the unified image–text sequence to represent shared content, text-unique information, image-unique information, cross-modal relations, and pragmatic analysis; a decision token at the end of the sequence integrates these states. A frozen decoder reconstructs the corresponding annotated field from each group. A frozen embedding teacher supplies the similarity distribution over other records, and supervised contrastive learning uses the categorical attribute labels. The final decision is made against three frozen semantic prototypes (*Neither*, *Broad-only*, and *Strict*), and probability marginalization ensures that the Strict probability never exceeds the Broad probability.

At inference time, a single image–text prefill produces both sarcasm and fine-grained predictions. For case analysis, the latent tokens can also be decoded into text to inspect the semantic information encoded for an example. Under unified training, SarcAtlas achieves the best result for every *Strict*/*Broad* combination across the four test sets. In cross-dataset evaluation, it leads all baselines in both *Strict* and *Broad* AUROC in every transfer direction, showing stronger transfer under the unified evaluation criteria.

Our contributions are threefold:

- **Unified annotation and analysis.** We introduce SarcAtlas, a unified annotation layer that explicitly characterizes sarcasm boundaries, mechanisms, and evidence. We use it to analyze systematic differences among existing multimodal sarcasm datasets.
- **Fine-grained annotation-guided latent-token modeling.** We supervise latent tokens for shared content, modality-specific information, pragmatic context, and image–text relations with their corresponding annotation fields. A decision token integrates these states and uses three frozen prototypes for hierarchically consistent Strict/Broad predictions.
- **Verifiable and efficient classification-style inference.** SarcAtlas achieves leading results in joint-training and cross-dataset generalization. One prefill produces predictions in parallel, while semantic decoding can be used for case studies to verify the decoded content against the original image and text.

2 SARCATLAS: FINE-GRAINED SARCASM DATA

Existing multimodal sarcasm datasets provide binary labels but apply different criteria to canonical irony and neighboring phenomena. SarcAtlas adds a unified fine-grained annotation layer to MMSD, MMSD2.0, and the English and Chinese portions of SarcNet. We retain the original image–text pairs, labels, and splits, as well as the correspondence between overlapping MMSD and MMSD2.0 examples. Figure 2 shows the review process and compares label boundaries, pragmatic types, evidence sources, and interpretation dependencies across the four dataset views.

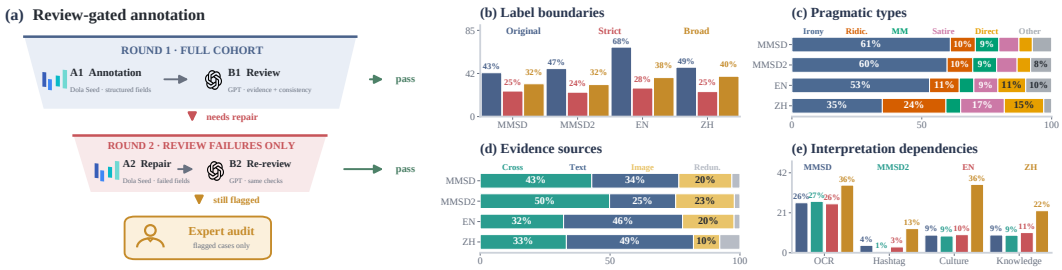


Figure 2: Overview of the reviewed annotations. (a) Web-search-enabled Dola Seed annotates and repairs, GPT reviews, and cases still flagged after re-review proceed to expert audit. (b–e) Dataset distributions of label boundaries, pragmatic types, evidence sources, and interpretation dependencies. Type and source distributions are conditioned on non-none annotations; dependency rates use all records. Values are rounded; small segments are unlabeled.

Annotation schema. Seven groups of fields separate the decision, its pragmatic mechanism, and its supporting evidence. **Decision boundaries** retain the Original label and add nested *Strict* and *Broad* judgments. Strict requires a recoverable implicit stance contrast or reversal; Broad also includes targeted, indirectly critical expressions without a canonical reversal. Direct criticism without an implicit sarcastic stance satisfies neither criterion. **Pragmatic mechanism** records type, target, and rationale independently of the boundary judgment. **Semantic content** distinguishes shared, text-unique, and image-unique information, while **cross-modal evidence** specifies the relation, primary source, and whether both modalities are required. The remaining groups record **affective structure**, **interpretation dependencies** such as OCR, hashtags, culture, and external knowledge, and **context and sensitivity**, including image genre and geographic or stakeholder dependence. Full definitions and boundary examples appear in Appendix B.1.

Annotation and review. In Figure 2(a), web-search-enabled Dola Seed generates a structured annotation and GPT reviews its evidential support, Strict/Broad hierarchy, and field completeness. Failed records receive field-targeted repair and the same review; those still flagged proceed to expert audit. Among 41,812 deduplicated annotation records, 86.5% passed the first review, 11.1% passed after repair, and the remaining 2.4% were finalized through expert audit. Further details on the annotation schema, model roles, and review procedure are provided in Appendix B.1.

Dataset analysis. Figure 2(b) shows that changes from Original labels to the unified boundaries occur in both directions: 34–44% of Original positives become Neither, 10–23% satisfy only Broad, and 1–15% of Original negatives satisfy Strict or Broad. This separates canonical sarcasm from neighboring cases while recovering examples that meet the shared criteria. Figure 2(c–e) shows further differences among datasets. Among records with a pragmatic type, verbal irony accounts for 61% in MMSD but 35% in SarcNet-ZH. Cross-modal evidence accounts for 50% of MMSD2.0 records with a non-none source but 33% in SarcNet-ZH; the latter relies more often on textual evidence and cultural context. These fields support separate analyses of decision boundaries, pragmatic mechanisms, and evidence requirements, rather than only an aggregate binary score.

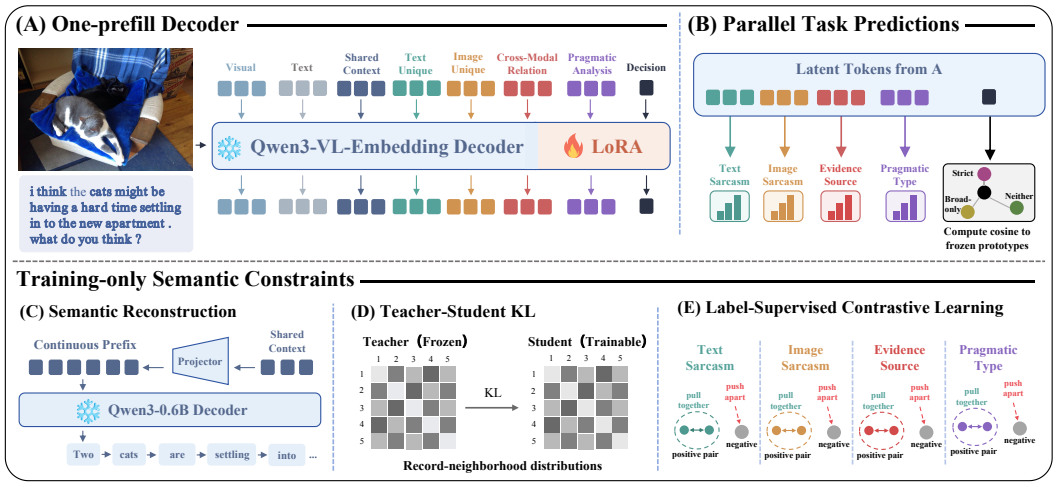


Figure 3: Overview of SarcAtlas. A single multimodal prefill contextualizes five latent-token groups and a decision token for parallel prediction. Training additionally uses semantic reconstruction, teacher–student neighborhood matching, and label-supervised contrastive learning.

3 METHOD

SarcAtlas is a single-tower model built on Qwen3-VL-Embedding-2B. Given an image–text pair, a single prefill jointly contextualizes the visual and text tokens, five groups of latent tokens, and one decision token. The five groups correspond to shared content, text-unique information, image-unique information, the cross-modal relation, and pragmatic analysis. The decision token aggregates this information for the nested *Strict*/*Broad* prediction. Figure 3 summarizes the model, its parallel prediction tasks, and the semantic constraints used during training.

3.1 LATENT-TOKEN MULTIMODAL ENCODING

As shown in Figure 3(A), let $x_n = (I_n, s_n)$ denote the n-th image–text pair. We append five latent-token groups and a decision token after its visual and textual tokens:

$$X_n = [V_n; W_n; E_c; E_t; E_i; E_r; E_p; e_d], \qquad H_n = F_{\theta_0,\Delta\theta}(X_n). \tag{1}$$

Here, V_n and W_n are the token sequences for image I_n and text s_n. The groups $E_c, E_t, E_i, E_r, E_p \in \mathbb{R}^{m\times d}$ correspond to shared content, text-unique information, image-unique information, the cross-modal relation, and pragmatic analysis, respectively; $e_d \in \mathbb{R}^d$ is the decision token. Each group contains $m = 3$ latent tokens, and d is the hidden dimension. The encoder $F_{\theta_0,\Delta\theta}$ is Qwen3-VL-Embedding-2B (Li et al., 2026), with frozen backbone parameters θ_0 and trainable LoRA parameters $\Delta\theta$ (Hu et al., 2021). All added tokens are initialized from the EOS embedding.

After the causal prefill, let $H_{n,k} \in \mathbb{R}^{m\times d}$ denote the contextualized representations of latent group k, and let $H_{n,d} \in \mathbb{R}^d$ denote the decision-token representation, where $k \in \mathcal{K} = \{c, t, i, r, p\}$. Semantic reconstruction uses all three representations in each group, whereas classification and contrastive learning use the group representation

$$z_{n,k} = \mathrm{LN}\left(\frac{1}{m}\sum_{j=1}^{m}\mathrm{LN}(H_{n,k,j})\right), \qquad z_{n,d} = \mathrm{LN}(H_{n,d}). \tag{2}$$

Here, $H_{n,k,j}$ is the contextualized representation of the j-th latent token in group k, and LN is LayerNorm without learnable affine parameters. Since the decision token is placed last in the causal sequence, $z_{n,d}$ integrates information from the image, text, and all five latent-token groups.

3.2 FINE-GRAINED PREDICTION AND HIERARCHICAL DECISION

Figure 3(B) shows the predictions computed in parallel from the same prefill. Four linear heads use $z_{n,t}$, $z_{n,i}$, $z_{n,r}$, and $z_{n,p}$ to predict text sarcasm, image sarcasm, evidence source, and pragmatic type, respectively. The shared-content representation $z_{n,c}$ has no separate attribute head.

The nested *Strict* and *Broad* boundaries define three mutually exclusive states, $\mathcal{C} = \{\text{NEITHER}, \text{BROAD-ONLY}, \text{STRICT}\}$. For each state, a frozen Qwen3-VL-Embedding model encodes a fixed set of bilingual definitions. We normalize the definition embeddings, average them within each state, and normalize the mean to obtain fixed semantic prototypes e_N, e_B, e_S. The cosine similarities between the decision representation and these prototypes define the state probabilities

$$q_{n,\ell} = \frac{\exp(\mathrm{sim}(z_{n,d}, e_\ell)/\tau_p)}{\sum_{\ell' \in \mathcal{C}} \exp(\mathrm{sim}(z_{n,d}, e_{\ell'})/\tau_p)}, \qquad \ell \in \mathcal{C}. \tag{3}$$

Here, $q_{n,\ell}$ is the probability of state ℓ, e_ℓ is its frozen prototype, sim denotes cosine similarity between normalized vectors, and τ_p is the prototype temperature. The two boundary probabilities are obtained by marginalizing the three-state distribution:

$$p_n^{\text{strict}} = q_{n,S}, \qquad p_n^{\text{broad}} = q_{n,B} + q_{n,S}. \tag{4}$$

Here, $q_{n,S}$ and $q_{n,B}$ are the probabilities of the STRICT and BROAD-ONLY states. This parameterization directly guarantees $p_n^{\text{strict}} \le p_n^{\text{broad}}$. The boundary and fine-grained predictions form the task loss

$$\mathcal{L}_{\text{task}} = \frac{1}{B} \sum_{n=1}^{B} \left[\mathrm{BCE}(p_n^{\text{strict}}, y_n^{\text{strict}}) + \lambda_b \, \mathrm{BCE}(p_n^{\text{broad}}, y_n^{\text{broad}}) + \lambda_a \sum_{a \in \mathcal{A}_n} \ell_{n,a} \right]. \tag{5}$$

Here, B is the batch size; y_n^{strict} and y_n^{broad} are the two boundary labels; $\mathcal{A}_n$ contains the available labels among the four attribute tasks; h_a is the linear head for task a; $k(a) \in \{t, i, r, p\}$ identifies its latent-token group; and y_n^a is the corresponding attribute label. We write $\ell_{n,a} = \mathrm{CE}(h_a(z_{n,k(a)}), y_n^a)$, while λ_b and λ_a weight the Broad boundary and each attribute task. Missing attribute labels do not contribute to the associated task loss. All loss weights are specified in Appendix C.3.

3.3 TRAINING-ONLY SEMANTIC CONSTRAINTS

The task loss directly supervises the predictions. The three objectives in Figure 3(C–E) additionally organize the latent tokens at the levels of field content, continuous semantic structure, and discrete attribute classes.

Semantic reconstruction. Figure 3(C) supervises the semantic content carried by each latent-token group. A shared projector P maps each of the three latent-token representations to two continuous prefix tokens, yielding six prefix tokens per group. A frozen Qwen3-0.6B decoder then generates the corresponding annotated field. The targets for groups c, t, i, r, p are shared content, text-unique information, image-unique information, the cross-modal relation, and the sarcasm explanation, respectively. The reconstruction loss is

$$\mathcal{L}_{\text{rec}} = \frac{1}{B|\mathcal{K}|} \sum_{n=1}^{B} \sum_{k \in \mathcal{K}} \overline{\mathrm{CE}}(D_{\phi_0}(P(H_{n,k})), R_{n,k}) \, . \tag{6}$$

Here, P is the shared prefix projector, D_{ϕ_0} is the Qwen3-0.6B decoder with frozen parameters ϕ_0, and $R_{n,k}$ is the annotated text for sample n and field k. The operator $\overline{\mathrm{CE}}$ denotes autoregressive cross-entropy averaged over the target tokens. The five fields contribute equally.

Teacher–student KL. Semantic reconstruction provides record-level field supervision. Figure 3(D) further distills the fine-grained similarity structure among the annotated texts into the corresponding latent-token space. For field k, the frozen Qwen3-VL-Embedding teacher encodes $R_{n,k}$ as $t_{n,k}$. The teacher representation $t_{n,k}$ and student representation $z_{n,k}$ query the same set of candidate teacher embeddings, producing two distributions over other records. Define

$$\Pi(v; \mathcal{T}, \tau) = \mathrm{softmax}(\mathrm{sim}(v, \mathcal{T})/\tau)\,, \tag{7}$$

where v is a query vector, $\mathcal{T}$ is a set of candidate teacher embeddings, τ is a temperature, and $\Pi(v; \mathcal{T}, \tau)$ is the resulting distribution of similarities to the candidate records. We minimize

$$\mathcal{L}_{\mathrm{nbr}} = \frac{1}{B|\mathcal{K}|} \sum_{n=1}^{B} \sum_{k \in \mathcal{K}} D_{\mathrm{KL}}\big(\Pi(t_{n,k}; \mathcal{T}_{\mathcal{Q}}^{k}, \tau_T) \,\big\|\, \Pi(z_{n,k}; \mathcal{T}_{\mathcal{Q}}^{k}, \tau_S)\big)\,. \tag{8}$$

Here, $\mathcal{T}_{\mathcal{Q}}^{k}$ contains the teacher embeddings for field k from the current batch and the teacher queue; τ_T and τ_S are the teacher and student temperatures; and D_{KL} is KL divergence. The current record's own teacher embedding is masked from both distributions. This objective transfers graded similarities between annotated fields, preserving fine-grained semantic structure within an attribute class.

Label-supervised contrastive learning. Figure 3(E) uses discrete attribute labels to shape the class structure of the associated latent-token groups. Text sarcasm, image sarcasm, evidence source, and pragmatic type supervise $z_{n,t}$, $z_{n,i}$, $z_{n,r}$, and $z_{n,p}$, respectively. Within each task, records with the same label form positive pairs, and all other valid candidates participate in the normalized contrastive objective (Khosla et al., 2020). Candidates come from the current batch and a detached feature queue. We average the four task-specific losses and denote the result by $\mathcal{L}_{\mathrm{con}}$. Teacher–student KL captures continuous similarities between annotated texts, whereas label-supervised contrastive learning imposes discrete class structure.

Combining the prediction and semantic objectives gives

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{nbr}}\mathcal{L}_{\mathrm{nbr}} + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}}. \tag{9}$$

Here, $\mathcal{L}_{\mathrm{task}}$ corresponds to the predictions in Figure 3(B), while $\mathcal{L}_{\mathrm{rec}}$, $\mathcal{L}_{\mathrm{nbr}}$, and $\mathcal{L}_{\mathrm{con}}$ correspond to Figure 3(C), (D), and (E), respectively. Their weights λ_{rec}, λ_{nbr}, and λ_{con} are given in Appendix C.3.

At inference, SarcAtlas retains the Qwen3-VL-Embedding backbone, LoRA adapters, four attribute heads, and three frozen prototypes. A single image–text prefill computes all predictions in parallel. For case studies, the frozen Qwen3-0.6B decoder can be loaded on demand to decode the five latent-token groups and verify whether the recovered semantic information is consistent with the sarcasm decision.

4 Experiments

We first evaluate the complete model's detection and transfer performance, then examine three intended properties: whether fine-grained supervision supports attribute prediction and structured latent spaces; whether the latent groups make intermediate semantic information inspectable; and whether the training-only semantic constraints preserve classification-style inference. All experiments retain the official splits. We report Strict/Broad Macro-F1 in-domain and Strict/Broad AUROC for direct transfer, select models only on source validation data, and provide the complete protocol in Appendix C.

4.1 Main Results under Unified Training

Every method is trained on the unified source in Table C1 and evaluated on all four native test sets; Table 1 reports Strict/Broad Macro-F1.

Table 1: Main results under deduplicated joint training. Each entry is Strict/Broad Macro-F1 (%). Best and second-best learned-model results are shown in bold and underlined, respectively. The final row reports absolute gains over the strongest baseline. Across the four test sets, the average SarcAtlas gain over Multi-View CLIP is $+4.50$ (95% CI $[2.25, 6.80]$) for Strict and $+3.24$ (95% CI $[1.25, 5.25]$) for Broad.

Method	MMSD	MMSD2.0	SarcNet-EN	SarcNet-ZH
Always negative	43.05 / 40.10	43.53 / 41.09	41.32 / 36.43	42.01 / 37.36
Always positive	19.62 / 24.84	18.64 / 23.23	22.83 / 29.92	21.61 / 28.75
mBERT (text only)	64.22 / 64.99	63.84 / 64.10	62.02 / 62.12	59.21 / 63.41
ViT (image only)	63.37 / 66.15	63.14 / 64.92	49.69 / 58.05	51.88 / 56.43
Frozen CLIP fusion	74.78 / 77.02	74.33 / 75.90	68.95 / 75.97	62.36 / 62.32
Frozen Qwen3-VL-Emb	73.68 / 75.40	71.65 / 74.36	66.47 / 70.77	68.50 / 70.20
HFM	71.09 / 72.20	68.48 / 69.65	66.99 / 71.95	70.74 / 68.79
Multi-View CLIP	76.87 / 80.26	76.81 / 79.56	68.31 / 75.46	71.68 / 70.15
DT4MID (XLM-R)	75.30 / 75.65	73.88 / 74.12	71.47 / 75.46	71.52 / 72.25
RCLMuFN	76.48 / 78.24	75.92 / 76.77	69.77 / 76.75	72.76 / 68.79
SarcAtlas (ours)	**82.52 / 80.86**	**79.97 / 80.14**	**75.19 / 80.72**	**74.00 / 76.65**
Δ *vs. best*	*+5.65 / +0.60*	*+3.16 / +0.58*	*+3.72 / +3.97*	*+1.24 / +4.40*

SarcAtlas ranks first on all eight dataset–boundary combinations, exceeding the strongest baseline by 1.24–5.65 Strict and 0.58–4.40 Broad points. Its four-dataset average improves over frozen Qwen3-VL-Embedding by 7.85/6.91 points; simultaneous gains across languages and boundaries rule out a gain obtained by trading off one setting against another.

4.2 Source-Specific In-Domain Results

To separate model design from the scale of unified training, we repeat the comparison with each source independently; SarcNet EN and ZH share one bilingual source but retain native test sets.

Table 2: Source-specific in-domain results. Each entry is Strict/Broad Macro-F1 (%). Best and second-best results are shown in bold and underlined, respectively. The final row reports absolute gains over the strongest baseline. Across the four matched source–test settings, the average SarcAtlas gain over Multi-View CLIP is $+3.74$ (95% CI $[1.59, 5.91]$) for Strict and $+2.62$ (95% CI $[0.60, 4.68]$) for Broad.

	Train: MMSD	**Train: MMSD2.0**	**Train: SarcNet EN+ZH**	
Method	MMSD	MMSD2.0	EN	ZH
mBERT	64.81 / 66.03	63.37 / 64.60	68.15 / 72.62	65.85 / 71.26
ViT	63.11 / 66.29	62.97 / 65.11	54.20 / 65.02	60.88 / 62.11
CLIP	75.67 / 78.14	74.33 / 75.25	62.58 / 71.97	67.46 / 69.25
Frozen Qwen3-VL-Emb	74.34 / 74.41	71.66 / 74.51	64.52 / 74.44	71.22 / **77.15**
HFM	70.09 / 70.36	66.27 / 66.46	59.16 / 62.31	66.17 / 63.48
MV-CLIP	77.85 / 78.87	75.52 / 77.85	68.40 / 74.34	72.64 / 71.33
DT4MID	73.75 / 74.52	72.24 / 72.80	65.91 / 73.22	70.44 / 74.68
RCLMuFN	76.80 / 77.85	76.46 / 78.16	65.45 / 73.02	69.46 / 68.51
SarcAtlas (ours)	**81.65 / 80.68**	**79.65 / 78.82**	**73.60 / 76.74**	**74.48** / 76.61
Δ *vs. best*	*+3.80 / +1.81*	*+3.19 / +0.66*	*+5.20 / +2.30*	*+1.84 / -0.54*

SarcAtlas leads all Strict and three of four Broad evaluations, so its gains do not depend on multi-source training. On SarcNet-ZH Broad it is close to Frozen Qwen3-VL-Emb (76.61 vs. 77.15), but

halves non-sarcasm false positives (61 to 30) and retains a 1.84-point Strict advantage, indicating a more selective Broad boundary without sacrificing the stricter criterion.

4.3 CROSS-DATASET TRANSFER

We apply each source-trained checkpoint directly to every off-source target, without fitting a target threshold; Table 3 reports Strict/Broad AUROC. MMSD-to-MMSD2.0 changes paired text and annotation policy, whereas SarcNet additionally changes content and, for ZH, language.

Table 3: Cross-dataset transfer in Strict/Broad AUROC (%). M1 and M2 denote MMSD and MMSD2.0. Best and second-best results are bold and underlined; the final row reports absolute gains over the strongest complete baseline. Across the eight transfer directions, the average SarcAtlas gain is $+11.75$ (95% CI $[9.68, 13.88]$) over RCLMuFN for Strict and $+8.81$ (95% CI $[6.84, 10.80]$) over CLIP for Broad.

	Train: MMSD			**Train: MMSD2.0**			**Train: SarcNet EN+ZH**	
Method	M2	EN	ZH	M1	EN	ZH	M1	M2
mBERT	74.52 / 71.70	63.22 / 59.14	44.00 / 47.30	73.71 / 70.13	67.55 / 65.66	47.39 / 48.29	55.53 / 57.02	57.48 / 57.99
ViT	73.46 / 73.51	54.90 / 59.17	47.05 / 54.99	72.06 / 73.80	54.94 / 60.28	43.89 / 55.19	58.68 / 65.81	60.05 / 65.77
CLIP	86.78 / 86.43	78.25 / 83.43	49.09 / 58.20	86.29 / 86.92	75.79 / 82.09	49.18 / 57.03	66.49 / 68.94	67.85 / 69.77
Frozen Qwen3-VL-Emb	83.72 / 83.95	72.85 / 75.73	49.30 / 61.86	84.09 / 84.51	72.28 / 76.14	52.95 / 59.51	67.46 / 71.79	67.08 / 71.56
HFM	77.09 / 75.27	60.49 / 57.67	45.85 / 47.31	77.96 / 77.27	64.90 / 63.01	50.03 / 42.73	53.50 / 53.37	54.61 / 53.63
MV-CLIP	87.44 / 87.37	72.03 / 79.76	41.21 / 55.79	87.47 / 88.51	75.46 / 80.64	43.82 / 55.08	69.21 / 70.96	70.28 / 72.45
DT4MID	83.10 / 82.35	74.69 / 77.12	53.76 / 60.74	85.50 / 83.86	75.25 / 76.55	56.67 / 62.66	64.46 / 62.96	66.13 / 65.98
RCLMuFN	87.49 / 86.57	78.43 / 83.40	47.57 / 56.81	87.17 / 87.20	79.75 / 83.73	41.94 / 51.59	68.18 / 68.25	71.94 / 74.20
SarcAtlas (ours)	**89.08 / 88.93**	**87.17 / 87.44**	**73.44 / 74.79**	**90.67 / 89.52**	**86.54 / 86.78**	**70.23 / 74.81**	**79.56 / 80.20**	**79.81 / 80.81**
Δ *vs. best*	*+1.59/+1.56*	*+8.74/+4.01*	*+19.68/+12.93*	*+3.20/+1.01*	*+6.79/+3.05*	*+13.56/+12.15*	*+10.35/+8.41*	*+7.87/+6.61*

SarcAtlas leads both AUROCs in every direction, with its largest margin under the cross-lingual shift to SarcNet-ZH ($+19.68/+12.93$ Strict/Broad). The learned criteria therefore remain useful across text, domain, and language shifts.

Table 4: Inference efficiency at batch size 4. Latency covers the complete end-to-end inference path. For latency, throughput, and MFU, best and second-best results are shown in bold and underlined, respectively.

Method	Params (M)	Latency ↓ (ms/batch)	Throughput ↑ (samples/s)	MFU ↑ (%)
Frozen CLIP fusion	149.625	78.84	50.73	7.18
HFM	81.834	104.06	38.44	2.68
Multi-View CLIP	164.744	89.27	44.81	3.01
DT4MID	364.539	**77.76**	**51.44**	7.09
RCLMuFN	359.239	154.23	25.94	2.38
SarcAtlas (ours)	2,127.018	197.62	20.24	**46.97**

4.4 CASE-LEVEL FINE-GRAINED INTERPRETATION

Aggregate scores do not show what semantic information the model encodes for an individual decision. Figure 4 therefore examines four correctly classified sarcastic and non-sarcastic posts. Fine-grained heads report attribute-level predictions, while five decoded fields recover semantic information encoded in the latent groups, including shared context, modality-specific cues, pragmatic mechanism, and cross-modal relation. Together, they provide an inspectable account of the intermediate representations that can be checked against the original image and text.

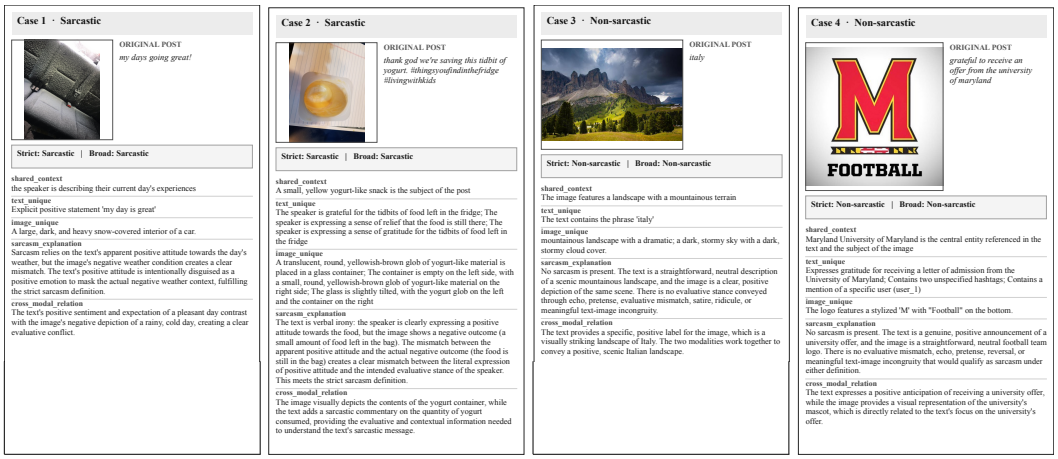


Figure 4: Four correct case studies with the original post, detection and attribute predictions, and five freely decoded semantic fields.

4.5 Fine-Grained Organization of Latent Spaces

Figure 5 shows attribute-specific organization rather than one generic sarcasm space. Text and image sarcasm occupy modality-specific regions; the relation space separates text-driven, image-driven, and cross-modal evidence; and the pragmatic space separates non-sarcastic cases before organizing sarcastic mechanisms. These geometries match the intended effect of groupwise contrastive supervision: each objective pulls together samples sharing the corresponding attribute without forcing incompatible semantic axes into one decision representation. The resulting specialization supports parallel prediction and case-level inspection.

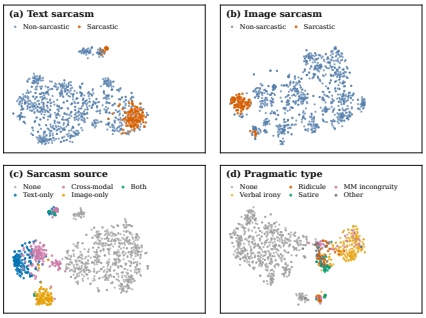


Figure 5: Attribute-specific t-SNE spaces for the same 1,000 validation examples. Colors denote the corresponding labels; 2,048-dimensional features are reduced to 50 dimensions by PCA before t-SNE.

4.6 Efficiency and Model FLOPs Utilization

At batch size 4 on one 32 GB GPU, we report deployed parameters, end-to-end latency, throughput, and model FLOPs utilization (MFU). Training-only modules are absent from the deployed graph; Table 4 summarizes the profile.

SarcAtlas reaches 20.24 samples/s and 46.97% MFU, 6.54× the next-highest MFU. Smaller models retain higher absolute throughput, but SarcAtlas uses the GPU more effectively for its richer workload. Thus, semantic objectives add training-time constraints without forcing autoregressive explanation into the classification path.

5 Conclusion

We introduced SarcAtlas, a unified fine-grained annotation layer for four English and Chinese dataset views, together with a single-tower model that assigns distinct latent-token groups to complementary semantic factors and enforces hierarchically consistent *Strict*/*Broad* decisions. SarcAtlas improves unified and source-specific detection and transfers more reliably across datasets and languages. Its case-level decodings and attribute-specific spaces make intermediate semantic evidence available for inspection, while a single image–text prefill retains classification-style inference. SarcAtlas thus provides both common decision boundaries and inspectable semantic representations for multimodal sarcasm judgments.

AI USE STATEMENT

We used generative AI tools (Dola Seed 2.1, GPT 5.5) to generate and review synthetic annotations such as labels and explanations, and to assist in language polishing and editing. AI-generated candidate annotations were randomly selected by the authors for review and revision against the source images, text, and annotation guidelines. All AI-edited text was reviewed and revised by the authors, who are responsible for the final content of the paper.

ETHICS STATEMENT

Multimodal sarcasm data may contain offensive language, political content, culturally specific references, identifiable individuals, and disagreements about intended meaning. We preserve source provenance and annotation uncertainty, analyze culture and stakeholder dependence, and do not redistribute raw media unless its license permits redistribution.

REPRODUCIBILITY STATEMENT

The main paper specifies the evaluated models, datasets and data splits, evaluation metrics, checkpoint-selection rules, and principal comparisons. The Appendix further provides the annotation protocol and field definitions, label and diagnostic definitions, result-aggregation procedures, training configurations, random seeds, cross-dataset evaluation settings, and representative cases to support verification and reproduction of the experimental process and reported results.

REFERENCES

Ibrahim Abu Farha, Silviu Vlad Oprea, Steven Wilson, and Walid Magdy. SemEval-2022 task 6: iSarcasmEval, intended sarcasm detection in English and Arabic. In Guy Emerson, Natalie Schluter, Gabriel Stanovsky, Ritesh Kumar, Alexis Palmer, Nathan Schneider, Siddharth Singh, and Shyam Ratan (eds.), *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 802–814, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.semeval-1.111. URL https://aclanthology.org/2022.semeval-1.111/.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

ByteDance Seed. Seed2.1: A next-generation agent for real-world productivity. Official model overview, 2026. URL https://seed.bytedance.com/en/seed2_1. Accessed 2026-09-26.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2506–2515, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1239. URL https://aclanthology.org/P19-1239/.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. Sentiment and emotion help sarcasm? a multi-task learning framework for multi-modal sarcasm, sentiment and emotion analysis. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4351–4360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.401. URL https://aclanthology.org/2020.acl-main.401/.

Zixin Chen, Hongzhan Lin, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen. CofiPara: A coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9663–9687, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.522. URL https://aclanthology.org/2024.acl-long.522/.

Herbert H. Clark and Richard J. Gerrig. On the pretense theory of irony. *Journal of Experimental Psychology: General*, 113(1):121–126, 1984. doi: 10.1037/0096-3445.113.1.121. URL https://doi.org/10.1037/0096-3445.113.1.121.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10563–10571, 2022. doi: 10.1609/aaai.v36i10.21300. URL https://ojs.aaai.org/index.php/AAAI/article/view/21300.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

Mengzhao Jia, Can Xie, and Liqiang Jing. Debiasing multimodal sarcasm detection with contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18354–18362, 2024. doi: 10.1609/aaai.v38i16.29795. URL https://ojs.aaai.org/index.php/AAAI/article/view/29795.

Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. Multi-source semantic graph-based multimodal sarcasm explanation generation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11349–11361, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.635. URL https://aclanthology.org/2023.acl-long.635/.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. E-ViL: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1244–1254, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Kayser_E-ViL_A_Dataset_and_Benchmark_for_Natural_Language_Explanations_in_ICCV_2021_paper.html.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5956–5968, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.411. URL https://aclanthology.org/2022.acl-long.411/.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause,

Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, July 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720, 2026. URL https://arxiv.org/abs/2601.04720.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4707–4715. Association for Computing Machinery, 2021. doi: 10.1145/3474085.3475190. URL https://doi.org/10.1145/3474085.3475190.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. Multi-modal sarcasm detection via cross-modal graph convolutional network. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1767–1777, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.124. URL https://aclanthology.org/2022.acl-long.124/.

Hui Liu, Wenya Wang, and Haoliang Li. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4995–5006, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.333. URL https://aclanthology.org/2022.emnlp-main.333/.

Yaochen Liu, Yazhou Zhang, Qiuchi Li, Benyou Wang, and Dawei Song. What does your smile mean? jointly detecting multi-modal sarcasm and sentiment using quantum probability. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 871–880, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.74. URL https://aclanthology.org/2021.findings-emnlp.74/.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

Silviu Oprea and Walid Magdy. iSarcasm: A dataset of intended sarcasm. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1279–1289, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.118. URL https://aclanthology.org/2020.acl-main.118/.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1383–1392, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.124. URL https://aclanthology.org/2020.findings-emnlp.124/.

Shraman Pramanick, Aniket Roy, and Vishal M. Patel. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 546–556, January 2022. doi: 10.1109/WACV51458.2022.00062. URL https://openaccess.thecvf.com/content/WACV2022/html/Pramanick_Multimodal_Learning_Using_Optimal_Transport_for_Sarcasm_and_Humor_Detection_WACV_2022_paper.html.

Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9507–9515, 2023. doi: 10.1609/aaai.v37i8.26138. URL https://ojs.aaai.org/index.php/AAAI/article/view/26138.

Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. MMSD2.0: Towards a reliable multi-modal sarcasm detection system. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10834–10845, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.689. URL https://aclanthology.org/2023.findings-acl.689/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. A multimodal corpus for emotion recognition in sarcasm. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6992–7003, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.756/.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 704–714, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1066/.

Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. NLX-GPT: A model for natural language explanations in vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8312–8322, June 2022. doi: 10.1109/CVPR52688.2022.00814. URL https://doi.org/10.1109/CVPR52688.2022.00814.

Rossano Schifanella, Paloma de Juan, Joel Tetreault, and LiangLiang Cao. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM International Conference on Multimedia*, pp. 1136–1145. Association for Computing Machinery, 2016. doi: 10.1145/2964284.2964321. URL https://doi.org/10.1145/2964284.2964321.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1732–1742, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.97. URL https://aclanthology.org/2024.naacl-long.97/.

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. Dynamic routing transformer network for multimodal sarcasm detection. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2468–2480, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.139. URL https://aclanthology.org/2023.acl-long.139/.

Divyank Tiwari, Diptesh Kanojia, Anupama Ray, Apoorva Nunna, and Pushpak Bhattacharyya. Predict and use: Harnessing predicted gaze to improve multimodal sarcasm detection. In Houda

Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15933–15948, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.988. URL https://aclanthology.org/2023.emnlp-main.988/.

David Tomás, Reynier Ortega-Bueno, Guobiao Zhang, Paolo Rosso, and Rossano Schifanella. Transformer-based models for multimodal irony detection. *Journal of Ambient Intelligence and Humanized Computing*, 14(6):7399–7410, 2023. doi: 10.1007/s12652-022-04447-y. URL https://doi.org/10.1007/s12652-022-04447-y.

Jiquan Wang, Lin Sun, Yi Liu, Meizhi Shao, and Zengwei Zheng. Multimodal sarcasm target identification in tweets. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8164–8175, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.562. URL https://aclanthology.org/2022.acl-long.562/.

Tongguan Wang, Junkai Li, Guixin Su, Yongcheng Zhang, Dongyu Su, Yuxue Hu, and Ying Sha. RCLMuFN: Relational context learning and multiplex fusion network for multimodal sarcasm detection. *Knowledge-Based Systems*, 319:113614, 2025. doi: 10.1016/j.knosys.2025.113614. URL https://doi.org/10.1016/j.knosys.2025.113614.

Deirdre Wilson. The pragmatics of verbal irony: Echo or pretence? *Lingua*, 116(10):1722–1743, 2006. doi: 10.1016/j.lingua.2006.05.001. URL https://doi.org/10.1016/j.lingua.2006.05.001.

Nan Xu, Zhixiong Zeng, and Wenji Mao. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3777–3786, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.349. URL https://aclanthology.org/2020.acl-main.349/.

Tan Yue, Xuzhao Shi, Rui Mao, Zonghai Hu, and Erik Cambria. SarcNet: A multilingual multimodal sarcasm detection dataset. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14325–14335, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.1248/.

Jiecheng Zhang, C.L.Philip Chen, Shuzhen Li, and Tong Zhang. Incongruity-aware tension field network for multi-modal sarcasm detection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14499–14508, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.705. URL https://aclanthology.org/2025.acl-long.705/.

A Related Work

A.1 Multimodal Sarcasm Detection

Multimodal sarcasm detection has largely been formulated as binary classification. Early work established the value of combining visual and textual evidence and introduced benchmarks such as MMSD (Schifanella et al., 2016; Cai et al., 2019). Subsequent methods improved fusion by modeling cross-modal contrast, intra- and inter-modal incongruity, relational structure, optimal transport, visual cues, and dynamic interaction (Xu et al., 2020; Pan et al., 2020; Liang et al., 2021; 2022; Liu et al., 2022; Pramanick et al., 2022; Qiao et al., 2023; Tian et al., 2023; Tomás et al., 2023; Tiwari et al., 2023). As lexical shortcuts and distribution shifts became more visible, recent work turned to debiasing, out-of-distribution evaluation, and normalized incongruity modeling (Jia et al., 2024; Tang et al., 2024; Zhang et al., 2025; Wang et al., 2025). MMSD2.0 and SarcNet further improved benchmark reliability through data cleaning, broader language coverage, and modality-specific annotation (Qin et al., 2023; Yue et al., 2024). These studies substantially advanced multimodal fusion and evaluation, but their primary supervision remains whether an example is sarcastic.

A.2 Fine-Grained Sarcasm Understanding

Beyond binary detection, prior work has introduced richer signals under different data and task settings. iSarcasm and iSarcasmEval use author-intended labels, ironic-speech categories, and non-sarcastic rephrases (Oprea & Magdy, 2020; Abu Farha et al., 2022). Conversational resources have been augmented with sentiment and emotion labels, and related models use affective prediction as auxiliary supervision (Chauhan et al., 2020; Liu et al., 2021; Ray et al., 2022); knowledge-enhanced detectors instead inject external information into the representation (Liu et al., 2022). Sarcasm target identification localizes who or what is targeted, whereas MuSE, MORE, and WITS provide or generate natural-language explanations (Wang et al., 2022; Chen et al., 2024; Desai et al., 2022; Kumar et al., 2022; Jing et al., 2023). These studies enrich particular datasets or modeling objectives, but the image–text benchmarks studied here lack a sample-aligned annotation that jointly connects the decision boundary, pragmatic mechanism, affect, target, evidence source, cross-modal relation, contextual dependence, and rationale.

A.3 Vision–Language Explanation and Representation

General vision–language research has progressed from dual-stream interaction and contrastive alignment to unified understanding–generation and large vision–language models (Lu et al., 2019; Radford et al., 2021; Li et al., 2022; Alayrac et al., 2022; Li et al., 2023). In particular, BLIP-2 uses a small set of learnable queries to compress and transfer information between a frozen vision encoder and a frozen language model. In parallel, natural-language explanation research has developed explanation benchmarks and unified generative models to make visual predictions more inspectable (Kayser et al., 2021; Sammani et al., 2022). These studies primarily target generic cross-modal representations or free-form explanations. We instead map multiple structured sarcasm semantics to latent tokens with designated roles, perform detection and fine-grained prediction in parallel after a single prefill, and decode the semantic tokens only for debugging and case analysis, making the model better suited to practical deployment.

B Annotation Protocol and Dataset Analysis

B.1 Operational Boundaries and Annotation Protocol

Strict and Broad views. We define Strict and Broad as progressively inclusive annotation criteria grounded in pragmatic accounts of irony and sarcasm. The **Strict** view identifies canonical sarcasm with an explicit pragmatic contrast: the surface stance and the recoverable evaluation are separated by distance, pretence, or reversal (Clark & Gerrig, 1984; Wilson, 2006; Riloff et al., 2013). The **Broad** view additionally includes social satire, ridicule, exaggeration, understatement, and rhetorical questions when they target a specific object and convey criticism indirectly, following the broader families used in intended-sarcasm annotation (Oprea & Magdy, 2020; Abu Farha et al., 2022). Pure

negative affect, literal criticism, ordinary humor, and image–text mismatch without an evaluative stance are negative under both views.

Structured record. Each record preserves the native image, text, split, identifier, and Original label, and organizes its annotations into seven field groups. The decision layer records modality-specific and boundary-specific judgments; the pragmatic layer describes mechanism and target; the content layer separates shared and modality-unique information; the relation layer specifies how image and text support the decision; the affective layer records modality-level sentiment; the dependency layer captures interpretation conditions such as OCR, hashtags, cultural context, and external knowledge; and the context-and-sensitivity layer records image genre, geographic scope, and stakeholder dependence. Table B1 summarizes the schema. Every record instantiates all required fields. Semantic absence or inapplicability is represented explicitly by values such as `none`, `false`, or `not applicable`, rather than by a missing label.

Table B1: **SarcAtlas annotation schema.** Natural-language descriptions, supporting evidence, and field-level rationales are stored alongside the structured labels summarized below.

Semantic layer	Main fields	Form	Purpose
Decision boundaries	Original label; text/image sarcasm; Strict; Broad	Binary	Preserve the source label and distinguish unimodal, Strict, and Broad multimodal decisions.
Pragmatic mechanism	Primary/secondary type; target; sarcasm explanation	Categorical, list, text	Identify the pragmatic mechanism, evaluated target, and reason that sarcasm does or does not hold.
Semantic content	Shared context; text-unique; image-unique	Text lists	Separate jointly expressed content from information contributed by only one modality.
Cross-modal evidence	Relation; evidence source; requires both; text spans/image regions	Categorical, binary, text	Describe the image–text relation, primary source of evidence, and whether both modalities are needed.
Affective structure	Text/image/joint polarity; emotion alignment	Categorical	Record modality-level and joint affect together with consistency, complementarity, or conflict across modalities.
Interpretation dependencies	OCR; hashtag; culture; external knowledge	Ordinal categories and evidence	Distinguish irrelevant, auxiliary, direct, helpful, and required cues, together with their supporting evidence.
Context and sensitivity	Image genre; country; continent; geographic scope; stakeholder sensitivity	Categorical	Characterize the visual form, geographic applicability, and potential stakeholder-dependent interpretation.

Evidence source takes one of `none`, `text`, `image`, `both-redundant`, `cross-modal`, and `uncertain`. OCR dependence uses `none`/`auxiliary`/`direct`; hashtag dependence additionally distinguishes `irrelevant`; and cultural and external-knowledge dependence use `none`/`helpful`/`required`. Emotion polarity is negative, neutral, positive, or mixed, while emotion alignment is consistent, complementary, contrasting, or unclear.

MMSD2.0 was developed from MMSD and retains the same 24,635 sample identities. For the 13,842 pairs whose native text or Original label changes, SarcAtlas stores an MMSD2.0-specific reviewed annotation; the remaining 10,793 unchanged pairs reuse the corresponding reviewed MMSD annotation. Consequently, 41,812 denotes the number of stored annotation records after deduplication, whereas the four reconstructed dataset views contain 52,605 records in total. All MMSD2.0 distributions reported below are computed over its complete 24,635-record view. We retain the official train, validation, and test partitions of every source dataset without resplitting. Identical MMSD

and MMSD2.0 pairs share an annotation record while retaining the native dataset view and official split metadata.

Annotation and review. The annotation task jointly requires visual understanding, OCR, structured reasoning, and access to cultural or event-specific background knowledge. We therefore select web-search-enabled Dola Seed 2.1 Turbo as the primary annotator. The official Seed2.1 model card reports scores of 90.1 on MathVision (92.7 with tools) and 80.1 on MMMU-Pro (82.2 with tools); the same evaluation reports 89.2 and 80.5, respectively, for Gemini 3.1 Pro. This comparison provides concrete evidence for the visual reasoning and tool-use capabilities required by our annotation protocol (ByteDance Seed, 2026). It produces complete structured annotations and performs field-targeted repair. GPT-5.5 independently reviews evidence support, label hierarchy, and field completeness, providing a separate model for quality control. Annotations that remain flagged after re-review proceed to expert audit. Of the 41,812 stored annotation records, 36,178 (86.5%) passed the first review, 4,644 (11.1%) passed after targeted repair and re-review, and the remaining 990 (2.4%) were manually audited and finalized. Across annotation, review, repair, and adjudication, the complete campaign consumed approximately **1.13 billion tokens**.

B.2 Golden Dataset Coverage and Distributions

The SarcAtlas registry contains 41,812 deduplicated annotation records. MMSD and MMSD2.0 contain partially overlapping examples; identical image–text pairs share one annotation record across the two dataset views. Table B2 reports statistics over the complete view of each dataset. Strict, Broad-only, and Neither form an exhaustive partition. The difference between Original and Strict prevalence ranges from 18.0 to 40.3 percentage points, while Broad-only cases occupy 7.2–15.1% of each dataset. These shifts show that the new views reorganize both positive examples and neighboring phenomena rather than merely relabeling a small set of outliers.

Table B2: **Dataset coverage and boundary distribution.** Counts are followed by percentages of all records in parentheses. Strict, Broad-only, and Neither form an exhaustive partition; no hierarchy violations remain.

Dataset	N	Original +	Strict	Broad-only	Neither
MMSD	24,635	10,647 (43.2%)	6,201 (25.2%)	1,764 (7.2%)	16,670 (67.7%)
MMSD2.0	24,635	11,655 (47.3%)	5,920 (24.0%)	1,853 (7.5%)	16,862 (68.4%)
SarcNet-EN	1,272	871 (68.5%)	358 (28.1%)	131 (10.3%)	783 (61.6%)
SarcNet-ZH	2,063	1,004 (48.7%)	508 (24.6%)	312 (15.1%)	1,243 (60.3%)

Aggregate prevalence does not reveal the direction of individual label changes. Table B3 therefore cross-tabulates each Original label with the Strict, Broad-only, and Neither decisions. Percentages are conditioned on the corresponding Original-positive or Original-negative row.

Table B3: **Original-to-boundary transition matrix.** Each cell reports the number of records followed by its percentage within the corresponding Original-label row.

Dataset	Original	N	Strict	Broad-only	Neither
MMSD	+	10,647	5,489 (51.6%)	1,065 (10.0%)	4,093 (38.4%)
	−	13,988	712 (5.1%)	699 (5.0%)	12,577 (89.9%)
MMSD2.0	+	11,655	5,412 (46.4%)	1,450 (12.4%)	4,793 (41.1%)
	−	12,980	508 (3.9%)	403 (3.1%)	12,069 (93.0%)
SarcNet-EN	+	871	354 (40.6%)	131 (15.0%)	386 (44.3%)
	−	401	4 (1.0%)	0 (0.0%)	397 (99.0%)
SarcNet-ZH	+	1,004	433 (43.1%)	230 (22.9%)	341 (34.0%)
	−	1,059	75 (7.1%)	82 (7.7%)	902 (85.2%)

The transitions are bidirectional: 34.0–44.3% of Original positives move to Neither, whereas 1.0–14.8% of Original negatives are recovered as Strict or Broad-only. A further 10.0–22.9% of Original

positives satisfy only the Broad criterion. The annotations therefore correct both overextended positives and missed sarcastic cases while separating canonical sarcasm from neighboring pragmatic phenomena.

Pragmatic type and the Strict/Broad boundary provide complementary views of an example: the same mechanism can fall on different sides of the boundary depending on how stance, target, and indirectness are realized in context. Table B4 shows that verbal irony is dominant in the three English datasets, whereas SarcNet-ZH contains substantially larger shares of ridicule, satire, and direct criticism. The remaining types combine exaggeration, punning, situational irony, understatement, rhetorical questions, and infrequent mechanisms.

Table B4: **Pragmatic-type distribution.** Percentages are conditioned on a non-none type annotation. "MM inc." denotes multimodal incongruity.

Dataset	Irony	Ridicule	MM inc.	Satire	Direct	Other
MMSD	61.0	9.6	9.0	7.7	5.2	7.4
MMSD2.0	59.8	9.9	9.2	7.9	5.2	8.1
SarcNet-EN	53.0	11.4	5.6	9.3	10.7	10.0
SarcNet-ZH	34.8	24.5	5.7	16.8	15.2	3.0

Evidence source and interpretation dependencies add analytical dimensions beyond the final label (Table B5). Cross-modal evidence accounts for 32.1–49.8% of records with a non-none source, and textual evidence is especially prominent in SarcNet. OCR is relevant to more than one quarter of every dataset, while the Chinese subset has the highest rates of hashtag, cultural, and external-knowledge dependence. The fields therefore identify distinct evaluation strata for subgroup analysis and intervention.

Table B5: **Evidence sources and interpretation dependencies (%).** Source percentages are conditioned on a non-none source; dependency rates use all records. "Redun." denotes evidence available redundantly in both modalities.

	Primary evidence source				Interpretation dependency			
Dataset	Cross	Text	Image	Redun.	OCR	Hashtag	Culture	Knowledge
MMSD	43.0	33.5	20.4	3.1	26.3	3.9	9.3	9.3
MMSD2.0	49.8	25.4	22.6	2.2	27.0	0.7	8.8	9.2
SarcNet-EN	32.1	45.8	19.8	2.2	25.7	3.1	9.5	10.6
SarcNet-ZH	33.2	48.9	10.1	7.8	35.5	12.7	35.9	22.1

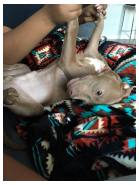
B.3 Case Studies

The following cases preserve native inputs and Original labels, then expose the structured evidence that determines the Strict and Broad decisions. They illustrate how the annotations correct dataset-specific boundaries and add diagnostic information beyond a final class label.

B.3.1 Original-Label Corrections and Boundary Shifts

Table B6 shows that binary labels conflate pragmatic content with dataset-specific annotation policy. Identical or minimally edited image–text pairs can cross the Original decision boundary even when their underlying pragmatic mechanism is unchanged. The same label space also admits literal criticism as positive while excluding canonical verbal irony in other cases. Our hierarchical annotation resolves this instability by placing attitude reversal in Strict, retaining neighboring ridicule, satire, and exaggeration in Broad, and excluding literal or metalinguistic uses from both. This normalization preserves informative boundary cases while making label disagreement interpretable as an annotation policy shift rather than an unexplained binary error.

Table B6: **Original-label conflicts and boundary corrections.** The left column preserves each native image–text input, the middle column reports the structured annotation, and the right column states what the case demonstrates.

Native case	Structured annotation	What this case demonstrates
(a) Strict cross-modal reversal  **Post (MMSD2).** my apartments told me that i can 't have my dog because he is a " vicious " breed . they right my dog is super vicious .	• **Original** – MMSD: **Non-sarcasm** – MMSD2: **Sarcasm** • **Ours** – Strict: **Sarcasm** – Broad: **Sarcasm** • **Type** Multimodal incongruity; verbal irony and ridicule • **Source** Cross-modal; requires both • **Relation** Contradiction • **Shared** The pictured dog is the speaker's dog • **Text-unique** The apartment bans the dog as a "vicious" breed; the speaker apparently agrees • **Image-unique** The dog lies on its back and appears calm and docile • **Strict basis** Apparent agreement with the apartment's claim is reversed by the image of a visibly non-threatening dog • **Broad basis** The contradiction ridicules the breed stereotype and the apartment's restriction	• **Policy inconsistency** The same pair and pragmatic mechanism receive opposite Original labels across MMSD and MMSD2 • **Controlled change** MMSD2 removes only `emoji_145`; this formatting change does not alter the cross-modal reversal • **Annotation gain** A unified Strict/Broad representation separates the stable sarcastic mechanism from the source annotation policy shift
(b) Broad-only political satire  **Post.** what a difference .	• **Original** – MMSD: **Non-sarcasm** – MMSD2: **Sarcasm** • **Ours** – Strict: **Non-sarcasm** – Broad: **Sarcasm** • **Type** Political satire; ridicule and exaggeration • **Source** Image-dominant; does not require both • **Relation** Complementary • **Shared** A marked difference between two public-speaking situations • **Text-unique** The contrast is notable • **Image-unique** Obama's "you/us/we" is contrasted with Trump's repeated "me" in two labeled political settings • **Strict basis** Absent: the post literally identifies the contrast, and the cartoon directly criticizes without performing a stance to reverse • **Broad basis** Present: the exaggerated comparison targets self-centered political rhetoric for ridicule	• **Policy inconsistency** Identical image and text receive opposite Original labels across the two dataset versions • **Boundary gain** The Broad label retains political ridicule, while the Strict label records that no attitude reversal is performed • **Dataset insight** A single binary label cannot distinguish rejection of the sample from disagreement over which neighboring pragmatic phenomena count as sarcasm

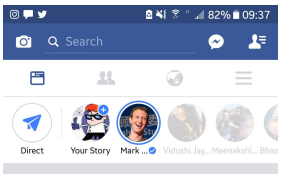
Continued on next page

Table B6 continued

Native case	Structured annotation	What this case demonstrates
(c) Direct criticism, not sarcasm  **Post.** not controlling air pollution is so dumb . this should be top priority [user]	• **Original** – MMSD2: **Sarcasm** • **Ours** – Strict: **Non-sarcasm** – Broad: **Non-sarcasm** • **Type** Direct criticism • **Source** None for sarcastic evidence; does not require both • **Relation** Complementary • **Shared** Air pollution is a serious concern requiring attention • **Text-unique** Failure to control pollution is foolish and should be a top priority • **Image-unique** A Delhi AQI dashboard reports 852, "Hazardous" • **Strict basis** Absent: the criticism is literal, with no pretense or attitude reversal • **Broad basis** Absent: the factual image supports the complaint without ridicule, satire, or ironic exaggeration	• **False positive** Strong negative sentiment and alarming visual evidence can be mistaken for sarcasm • **Neighbor boundary** Direct criticism is close to sarcasm in stance, but lacks its pragmatic mechanism • **Annotation gain** Strict/Broad labels preserve this distinction instead of treating every critical incongruity as sarcasm
(d) Mention, not use  **Post.** made me think of [friends] and [friends] miss you gals ! #inappropriatelydressed	• **Original** – MMSD: **Sarcasm** • **Ours** – Strict: **Non-sarcasm** – Broad: **Non-sarcasm** • **Type** Sarcasm mention; personality meme • **Source** None for sarcastic evidence; the image supplies the mention • **Relation** Complementary • **Shared** The post associates the pictured personality description with the author's friends • **Text-unique** The author thinks of and misses those friends • **Image-unique** A quote lists inappropriate comments, sarcasm, and random shenanigans as shared traits • **Strict basis** Absent: no apparent stance is reversed in the current utterance • **Broad basis** Absent: the sample names sarcasm but performs no ridicule, satire, or sarcastic act	• **Lexical leakage** The word "sarcasm" is visually present even though the sample itself is not sarcastic • **Noise audit** Keyword or meme-template shortcuts can produce a correct Original prediction for the wrong reason • **Annotation gain** The schema distinguishes talking about sarcasm from actually using it

Continued on next page

Table B6 continued

Native case	Structured annotation	What this case demonstrates
(e) Original positive, Broad-only  **Post.** when only founder believes in the product , and no one else .	• **Original** – MMSD2: **Sarcasm** • **Ours** – Strict: **Non-sarcasm** – Broad: **Sarcasm** • **Type** Ridicule; situational irony • **Source** Cross-modal; requires both • **Relation** Complementary • **Shared** Only a product's founder appears to support or use it • **Text-unique** The unnamed product is framed as unsuccessful • **Image-unique** Facebook Stories shows Mark Zuckerberg as the only user with an active story • **Strict basis** Absent: the statement directly presents its critical stance rather than performing the opposite one • **Broad basis** Present: the image grounds a clear target of ridicule	• **Positive-label ambiguity** An Original positive can encode ridicule without canonical attitude reversal • **Boundary gain** Strict and Broad retain the useful neighboring phenomenon while recording the missing reversal • **Semantic gain** Source and shared/unique fields show why both modalities are required to identify the target
(f) Missed Broad-only exaggeration  **Post.** people who leave drinks like this can not be trusted	• **Original** – MMSD2: **Non-sarcasm** • **Ours** – Strict: **Non-sarcasm** – Broad: **Sarcasm** • **Type** Exaggeration • **Source** Cross-modal; requires both • **Relation** Complementary • **Shared** The image depicts the "drinks like this" behavior named in the post • **Text-unique** Leaving a small remainder is treated as evidence that a person cannot be trusted • **Image-unique** A nearly empty bottle contains a small amount of orange liquid • **Strict basis** Absent: expressed and intended attitudes are both negative, differing only in severity • **Broad basis** Present: moral condemnation is deliberately disproportionate to a minor annoyance	• **False negative** Original Non-sarcasm can omit recognizable exaggerated ridicule • **Boundary gain** Broad captures disproportionate mockery without misrepresenting it as attitude reversal • **Evidence gain** The source fields record that the visual state of the bottle anchors the exaggeration

Continued on next page

Table B6 continued

Native case	Structured annotation	What this case demonstrates
(g) Missed Strict verbal irony  **Post.** blessed to announce that i am committed to acl surgery university and 6–9 months of recovery ! ! !	• **Original** – MMSD2: **Non-sarcasm** • **Ours** – Strict: **Sarcasm** – Broad: **Sarcasm** • **Type** Verbal irony • **Source** Text; does not require both • **Relation** Complementary • **Shared** The subject is a school basketball player • **Text-unique** A celebratory college-commitment formula is used for ACL surgery and a long recovery • **Image-unique** A posed basketball portrait supplies the athletic setting being parodied • **Strict basis** Present: positive commitment language pretends to celebrate an unwanted injury • **Broad basis** Present through the same verbal irony	• **False negative** A canonical verbal-irony construction is absent from the Original positive class • **Source insight** Text alone establishes the reversal; the image adds context but is not required • **Annotation gain** The unified labels correct the decision while the source field prevents an unsupported claim of multimodal necessity

B.3.2 What the Structured Fields Add

Table B7 demonstrates that the structured fields add decision-relevant information that cannot be recovered from the final label. Shared and modality-unique descriptions localize the semantic content, while relation and rationale fields specify how that content supports a pragmatic decision. The dependency annotations further distinguish necessary cues from corroborating or irrelevant ones: affective mismatch alone is not sufficient for sarcasm, OCR and hashtags play graded roles, and a multimodal record may be text-driven, image-driven, genuinely cross-modal, or evidentially redundant. Together with target, knowledge, geography, and image-genre metadata, these fields define intervention-ready subgroups and make it possible to attribute a model failure to a specific reasoning requirement rather than to generic multimodal difficulty.

Table B7: **Information added by the structured fields.** Each row retains the native image–text input. The middle column reports only the fields needed for the controlled comparison; the right column states the dataset-level conclusion supported by that comparison.

Native case	Structured annotation	What this case demonstrates
(a) Semantic decomposition  **Post.** another sensitive and considerate 1950s magazine advert aimed at women	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Shared** An advertisement aimed at women • **Text-unique** The ad is from the 1950s; "sensitive and considerate"; "another" frames it as a recurring pattern • **Image-unique** A Rallie slimming belt, its product claims, a woman's illustration, and body-shaming copy • **Relation** Sentiment mismatch: apparent praise conflicts with the ad's misogynistic framing	• **Non-redundancy** Shared and modality-unique fields preserve different facts rather than paraphrasing the binary decision • **Composition** The relation field records how those facts combine into an ironic interpretation • **Auditability** Each semantic claim can be checked against one or both input modalities

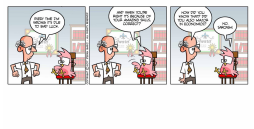
Continued on next page

Table B7 continued

Native case	Structured annotation	What this case demonstrates
(b) Boundary-aware explanation  **Post.** "except the burning diarrhoea i experienced that time i ate three activia in a row."	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Non-sarcasm**; Broad **Sarcasm** • **Type** Multimodal incongruity and ridicule • **Source** Cross-modal; requires both • **Relation** Contradiction • **Strict basis** The post literally reports a negative experience; it does not perform an opposite stance • **Broad basis** The Activia lid's "I AM thankful FOR EVERY MOMENT" message becomes a visual punchline that mocks the product promise	• **Explanation gain** A rationale can explain why a pair belongs to Broad sarcasm while failing the Strict criterion • **Boundary gain** The explanation is tied to two explicit decisions, instead of rationalizing one undifferentiated positive label
(c) Target and sensitivity  **Post.** rare photo of donald j. trump and steve bannon together.	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Non-sarcasm**; Broad **Sarcasm** • **Type** Political satire • **Target** The two named political figures and the ideology implied by their visual comparison • **Culture/knowledge** Required • **Stakeholder sensitivity** High	• **Target gain** A positive label alone does not identify who or what is being mocked • **Release safety** Sensitivity metadata enables stratified audits of political and historical comparisons
(d) Hashtag: direct  **Post.** [user] what is this bullcrap. now i have to go back to the store. #goodjob	• **Original**: MMSD and MMSD2 **Sarcasm** • **With hashtag**: Strict **Sarcasm**; Broad **Sarcasm** • **Without hashtag**: Strict **Non-sarcasm**; Broad **Non-sarcasm** • **Dependency** Direct • **Reason** #goodjob supplies the apparent praise that reverses the literal complaint about spoiled food	• **Controlled intervention** Removing one linguistic cue flips both boundary decisions while the image and remaining post are fixed • **Dataset revision** This makes hashtag removal a paired semantic shift, not ordinary cross-dataset transfer
(e) Hashtag: auxiliary  **Post.** feeling anxious already? drink a rockstar, that'll help, genius. #truestory #livingit	• **Original**: MMSD **Sarcasm**; MMSD2 **Non-sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Dependency** Auxiliary • **Reason** "that'll help, genius" already establishes verbal irony; the hashtags reinforce self-mocking, lived-experience framing • **After removal** Strict/Broad remain Sarcasm	• **Graded dependence** A removed hashtag can change tone without changing the pragmatic class • **Label diagnosis** The MMSD2 label flip cannot be assumed to follow from the hashtag removal itself

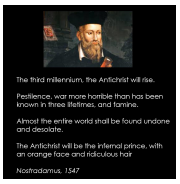
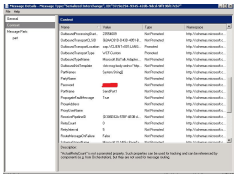
Continued on next page

Table B7 continued

Native case	Structured annotation	What this case demonstrates
(f) Hashtag: irrelevant  **Post.** i love deadlines. i love the whooshing noise they make as they go by. — douglas adams #writing #amwriting #film #books #art	• **Original**: MMSD2 **Non-sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Source** Text • **Dependency** Irrelevant • **Reason** The quote contains the full reversal; the hashtags only index content categories • **After removal** Strict/Broad remain Sarcasm	• **Negative control** Hashtag presence is not itself evidence of sarcasm • **Shortcut audit** Direct, auxiliary, and irrelevant cases support a three-level intervention test instead of a binary hashtag flag
(g) OCR: direct  **Post.** today's offshorecomic. skill vs luck.	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Source** Image • **OCR dependency** Direct • **Reason** The neutral post names the comic; its sarcastic dialogue and punchline are entirely inside the image	• **Visual-language requirement** Image understanding here means reading and relating embedded text, not only recognizing objects • **Deployment implication** OCR-sensitive performance should be reported separately from generic image dependence
(h) OCR: auxiliary  **Post.** congratulations boots for managing a massive 5% saving over the branded medicines.	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Source** Text • **OCR dependency** Auxiliary • **Reason** "massive 5% saving" already marks ironic exaggeration; the prices £3.15 and £2.99 only verify its factual basis	• **Functional distinction** OCR may establish the sarcastic mechanism or merely corroborate it • **Robustness test** Masking image text should affect direct cases more than auxiliary cases
(i) Emotion conflict: sarcastic  **Post.** what a nice guy. this attitude definitely doesn't help create an environment of fear and hatred which can lead to radicalization	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Emotion** Text positive; image mixed/negative; joint mixed • **Alignment** Contrasting • **Relation** Sentiment mismatch • **Reason** Apparent praise reverses the judgment of a violent threat shown in the image	• **Positive example** Affective mismatch contributes to sarcasm when it realizes an identifiable stance reversal • **Structured evidence** Emotion labels expose that mechanism without equating every cross-modal affective conflict with sarcasm

Continued on next page

Table B7 continued

Native case	Structured annotation	What this case demonstrates
(j) Emotion conflict: literal  **Post.** the thing that pissed me off the most was this. really [user]? you couldn't just let us watch the credits? you had to put an ad?	• **Original**: MMSD2 **Non-sarcasm** • **Ours**: Strict **Non-sarcasm**; Broad **Non-sarcasm** • **Type** Direct criticism • **Emotion** Negative complaint against a positive entertainment image • **Relation** Complementary • **Reason** The speaker literally objects to an advertisement covering the credits; no stance is pretended or reversed	• **Matched negative control** A cross-modal emotion conflict can be a literal complaint • **Field interaction** Relation, explanation, and boundary fields are needed to interpret emotion rather than using it as a label proxy
(k) Culture and external knowledge  **Post.** i think he might have been on to something	• **Original**: MMSD2 **Non-sarcasm** • **Ours**: Strict **Non-sarcasm**; Broad **Sarcasm** • **Type** Political satire • **Source** Image • **Culture** Required • **Knowledge** Required: Nostradamus's prophetic persona and the visual allusion to Donald Trump	• **Hidden condition** The image is interpretable only after resolving two culturally situated references • **Failure diagnosis** A wrong prediction can be localized to missing background knowledge rather than generic cross-modal fusion
(l) Geography: local  **Post.** way to go #schererville for keeping indiana in the news	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Scope** Local • **Country/continent** United States / North America • **Evidence** Schererville, Indiana and the local gun-related news shown in the image	• **Grounded locality** Geographic metadata is assigned from explicit place and event evidence • **Transfer analysis** Local-event cases can be separated from globally interpretable sarcasm
(m) Geography: global  **Post.** forgot the sap password. good thing biztalk stores it plain text in the context! which is nice . . .	• **Original**: MMSD2 **Non-sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Source** Text • **Scope** Global • **Country** Uncertain • **Knowledge** Helpful: enterprise-software security practice	• **Negative geographic control** English and product names do not justify assigning a country • **Schema gain** Scope and country are separated so globally shared technical context is not mislabeled as national culture

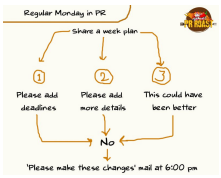
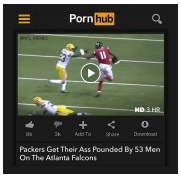
Continued on next page

Table B7 continued

Native case	Structured annotation	What this case demonstrates
(n) Evidence source: text  **Post.** i'm happy with this because we definitely don't need clean water & air. and our weather has been so normal	• **Original**: MMSD **Sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Source** Text • **Requires both** No • **Reason** The positive surface wording and intended environmental criticism coexist in the post itself	• **Text-source anchor** A multimodal record need not require the image for its sarcastic mechanism • **Intervention prediction** Image removal should preserve the core decision for this subgroup
(o) Evidence source: image  **Post.** presidents day special:	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Non-sarcasm**; Broad **Sarcasm** • **Source** Image • **Requires both** No • **Reason** The post is only a neutral title; visual content supplies the political ridicule	• **Image-source anchor** Sarcasm can be contained in an image even when a separate post accompanies it • **Diagnostic value** Source labels distinguish image understanding from cross-modal composition
(p) Evidence source: cross-modal  **Post.** well done	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Source** Cross-modal • **Requires both** Yes • **Relation** Sentiment mismatch • **Reason** Praise becomes a condemnation only when combined with the negative event in the image	• **Cross-modal anchor** Neither modality alone fixes the intended stance • **Three-way source taxonomy** Rows (n)–(p) isolate text, image, and cross-modal evidence with concrete counterexamples
(q) Both modalities required  **Post.** such a beautiful moment.	• **Original**: MMSD2 **Non-sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Source** Cross-modal • **Requires both** Yes • **Reason** "beautiful" reverses its valence only in relation to the negative visual event	• **Necessity** The second modality changes the interpretation rather than merely adding detail • **Intervention target** Removing either modality should destroy the recoverable reversal
(r) Both modalities redundant  **Post.** deleting tests is the only surefire way to make them pass.	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Source** Both redundant • **Requires both** No • **Reason** The same software-testing punchline is recoverable from the post and from the screenshot	• **Redundancy** Two modalities may repeat evidence without requiring cross-modal reasoning • **Shortcut audit** High multimodal accuracy on redundant samples is not evidence that a model composes modalities

Continued on next page

Table B7 continued

Native case	Structured annotation	What this case demonstrates
(s) Genre: real photo  **Post.** [user] this was bought at your store on saturday . . . [num] days in the fridge and it's already rotten!	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Non-sarcasm**; Broad **Non-sarcasm** • **Genre** Real photo • **Visual requirement** Object state and spoilage recognition	• **Genre-specific demand** Real photographs emphasize physical-state recognition • **Boundary control** Visually grounded dissatisfaction is not automatically sarcastic
(t) Genre: infographic/cartoon  **Post.** just a regular monday in pr . . .	• **Original**: MMSD2 **Sarcasm** • **Ours**: Strict **Sarcasm**; Broad **Sarcasm** • **Genre** Infographic/cartoon • **Visual requirement** OCR plus the relations encoded by a designed diagram	• **Genre-specific demand** Designed graphics require reading symbols, labels, and layout relations • **Evaluation gain** Genre permits performance breakdowns that an undifferentiated image flag cannot support
(u) Genre: meme  **Post.** if anyone wants the highlights, you can find them here . . .	• **Original**: MMSD **Non-sarcasm** • **Ours**: Strict **Non-sarcasm**; Broad **Sarcasm** • **Type** Pun humor • **Source** Image • **Genre** Meme/platform parody • **Visual requirement** Template recognition and sports/platform context	• **Genre-specific demand** Meme interpretation depends on a reusable visual convention, not only depicted objects • **Boundary gain** Genre and type jointly describe why the sample is Broad-only

C Experimental Protocol and Implementation Details

This section specifies the data partitions, input processing, model selection, free-decoding and visualization protocols, model configuration, comparison controls, evaluation metrics, and efficiency measurement used in our experiments.

C.1 Datasets and Splits

We evaluate the SarcAtlas views of MMSD, MMSD2.0, SarcNet-EN, and SarcNet-ZH. All four datasets use their official train, validation, and test membership, and models receive the native image and text provided by each source.

MMSD2.0 revises MMSD and shares its sample identities and split membership. This correspondence supports paired evaluation of text revisions and the source annotation policy shift under the same Strict/Broad definitions. The unified training source is organized by record version: records with identical content and labels share one training instance, whereas revised MMSD2.0 records retain their corresponding version. The resulting source contains 32,920 training and 4,423 validation instances. We report performance separately on the four official test sets.

Table C1: **Official split sizes used in our experiments.** The unified source supports joint training and validation; evaluation is reported on its four component test sets.

Dataset / training source	Language	Train	Validation	Test
MMSD	English	19,816	2,410	2,409
MMSD2.0	English	19,816	2,410	2,409
SarcNet-EN	English	758	247	267
SarcNet-ZH	Chinese	1,240	424	399
Unified source	English + Chinese	32,920	4,423	–

C.2 Input Processing, Model Selection, and Free Decoding

Models receive the original image and text supplied by each dataset, without manual rewriting or whitespace normalization. Images are processed with the native dynamic-resolution mechanism of Qwen3-VL, using minimum and maximum pixel counts of 3,136 and 200,704, respectively. We impose no additional text truncation at the backbone input. After assembling the multimodal input sequence, we append five groups of semantic readout tokens followed by one decision token. In the main experiments, the maximum training length of each of the five reconstruction targets is 96 tokens.

All models are trained for five epochs. We select the best checkpoint by Strict Macro-F1 on the validation set of the training source; joint training uses the unified validation set. The same selected checkpoint is then evaluated directly on all four test sets, and no test result enters model selection.

The five textual fields used for case analysis are obtained by label-free decoding. At inference, the model receives only the original image, its paired text, and the same readout prompts used during training; it receives no Golden label, reference explanation, or target field. Text is generated separately from the latent states associated with shared context, text-unique information, image-unique information, sarcasm explanation, and cross-modal relation. We use deterministic greedy decoding without sampling, allow at most 96 new tokens per field, and stop when the end-of-sequence token is produced. The four complete test sets contain 2,409, 2,409, 267, and 399 records, respectively, for 5,484 records in total. Case-study examples are selected only from completed full-test outputs for which all five fields are nonempty.

C.3 SarcAtlas Implementation

Model and latent states. SarcAtlas uses Qwen3-VL-Embedding-2B as its backbone. The official processor maps each image–text pair to one multimodal sequence. We append three trainable tokens for each of the five semantic roles and one decision token; all 16 tokens are initialized from the EOS

embedding. Their native 2048-dimensional hidden states are retained for reconstruction, while the three states within each semantic role are mean-pooled after layer normalization for classification and contrastive learning. The backbone is frozen except for rank-16 LoRA adapters ($\alpha = 16$, dropout $= 0.05$) on the query, key, value, gate, up, and down projections and the 16 new token embeddings.

Reconstruction and decision prototypes. Semantic reconstruction uses a frozen Qwen3-0.6B decoder. One shared layer-normalized linear projector maps every latent vector to two decoder prefix tokens, so each three-token group provides six prefixes. The five reconstruction fields receive equal weight after averaging token losses within each record. Fixed detection prototypes use four English and four Chinese descriptions for each of NEITHER, BROAD-ONLY, and STRICT. The decision state is compared directly with these 2048-dimensional prototypes at temperature 0.07.

Teacher–student KL. For each semantic field, the frozen Qwen3-VL-Embedding teacher and its corresponding student latent state query the same candidate teacher embeddings. The candidate set combines the current batch with a 256-entry teacher queue for that field. The current record's own teacher embedding is excluded from both similarity distributions. We use temperatures $\tau_T = \tau_S = 0.07$ for the teacher and student distributions. The teacher and its queue are used only during training.

Label-supervised contrastive learning. The text-sarcasm label supervises the text-unique state, the image-sarcasm label supervises the image-unique state, the evidence-source label supervises the relation state, and the pragmatic-type label supervises the pragmatic state. Let $\mathcal{A}$ denote these four tasks. For task a, $\mathcal{C}_a(n)$ contains all valid candidates for anchor n from the current batch and the corresponding detached feature queue, excluding the anchor itself. The positive set $\mathcal{P}_a(n) \subseteq \mathcal{C}_a(n)$ contains candidates with the same task label. With $\mathcal{I}_a$ denoting anchors that have at least one positive, the task-specific objective is

$$\mathcal{L}^a_{\text{con}} = -\frac{1}{|\mathcal{I}_a|} \sum_{n \in \mathcal{I}_a} \frac{1}{|\mathcal{P}_a(n)|} \sum_{p \in \mathcal{P}_a(n)} \log \frac{\exp\big(\text{sim}(z^a_n, z^a_p)/\tau_{\text{con}}\big)}{\sum_{q \in \mathcal{C}_a(n)} \exp\big(\text{sim}(z^a_n, z^a_q)/\tau_{\text{con}}\big)}. \tag{10}$$

Here, z^a_n is the normalized latent state for task a, sim is cosine similarity, and $\tau_{\text{con}} = 0.1$. We average $\mathcal{L}^a_{\text{con}}$ over the four tasks. Missing labels are encoded as -100 and excluded from the corresponding candidate set. Anchors without a positive are skipped; if a task has no valid positive pair, its loss is zero. Each task maintains a class-balanced feature queue of 512 entries. Queue entries are retained for at most eight optimizer updates, and all feature queues are cleared at each epoch boundary.

Optimization. Table C2 summarizes the final model and training configuration.

Table C2: **SarcAtlas model and training configuration.**

Configuration	Setting
Backbone	Qwen3-VL-Embedding-2B
Latent and decision tokens	Five groups of three latent tokens; one decision token; EOS initialization
LoRA	Rank 16; $\alpha = 16$; dropout 0.05; query, key, value, gate, up, and down projections
Reconstruction decoder	Frozen Qwen3-0.6B; two continuous prefix tokens per latent token
Detection prototypes	NEITHER, BROAD-ONLY, and STRICT; four English and four Chinese descriptions per prototype; temperature 0.07
Optimizer	AdamW; learning rate 3×10^{-4}; weight decay 0.01
Schedule	Five epochs; linear schedule; 10% warmup
Batch configuration	Micro-batch 16; four gradient-accumulation steps; effective batch size 64
Numerical configuration	BF16; maximum gradient norm 5
Prediction-loss weights	Strict 1.0; Broad 0.5; each attribute task 0.25
Semantic-loss weights	Reconstruction 0.3; teacher–student KL 2.0; label-supervised contrastive learning 0.1
Teacher–student temperatures	$\tau_T = 0.07$; $\tau_S = 0.07$
Contrastive temperature	$\tau_{\text{con}} = 0.1$
Random seed	42

In the notation of Section 3, these settings correspond to $\lambda_b = 0.5$, $\lambda_a = 0.25$, $\lambda_{\text{rec}} = 0.3$, $\lambda_{\text{nbr}} = 2.0$, and $\lambda_{\text{con}} = 0.1$.

C.4 Representation Visualization Protocol

Figure 5 visualizes the complete model's latent representations in four fine-grained semantic spaces: text sarcasm, image sarcasm, sarcasm evidence source, and pragmatic type. For the same 1,000 randomly sampled validation examples, each semantic state is taken as the raw 2,048-dimensional latent vector produced by the selected post-training checkpoint. We first apply L_2 normalization, reduce the vectors to 50 dimensions with PCA, and then project them to two dimensions with t-SNE using cosine distance. Each panel fits its projection independently. We set t-SNE perplexity to 30, use the automatically selected learning rate and PCA initialization, run 2,000 iterations, and fix the random seed to 42. Colors indicate the ground-truth class of the corresponding field and do not enter the dimensionality-reduction procedure.

C.5 Fine-Grained Attribute Prediction

To examine whether the model actually learns from fine-grained semantic supervision, we further report Macro-F1 for the four attribute heads of the jointly trained model on all four complete test sets: text sarcasm, image sarcasm, sarcasm evidence source, and pragmatic type. The results are shown in Table C3.

Table C3: Fine-grained attribute prediction by the jointly trained SarcAtlas on the four complete test sets. Values are Macro-F1 (%). Text, Image, Source, and Type denote text sarcasm, image sarcasm, sarcasm evidence source, and pragmatic type, respectively. Uniform random is the reference Macro-F1 implied by independent uniform prediction over the declared labels and the empirical class prevalences.

Method	MMSD				MMSD2.0				SarcNet-EN				SarcNet-ZH			
	Text	Image	Source	Type	Text	Image	Source	Type	Text	Image	Source	Type	Text	Image	Source	Type
Uniform random	41.36	39.04	11.17	4.03	39.30	38.80	10.76	3.91	44.72	38.73	11.82	4.67	45.25	37.66	11.64	4.83
SarcAtlas (joint)	**77.03**	**68.74**	**54.29**	**26.74**	**71.38**	**70.63**	**50.57**	**23.83**	**78.96**	**65.36**	**50.08**	**20.61**	**80.16**	**67.41**	**41.58**	**31.39**

SarcAtlas exceeds the uniform-random reference in all 16 dataset–attribute evaluations. The text-sarcasm head achieves Macro-F1 scores of 71.38–80.16 across the four test sets and is the most stable, indicating that the model can reliably recover text-unique sarcasm semantics from the jointly contextualized multimodal representation. The image-sarcasm head obtains 65.36–70.63, showing that the same representation also supports image-unique sarcasm semantics. Evidence source and pragmatic type, by comparison, require the model to further distinguish where the cross-modal conflict arises and which pragmatic mechanism it instantiates. These tasks contain more classes and finer decision boundaries, resulting in Macro-F1 ranges of 41.58–54.29 and 20.61–31.39, respectively. Overall, the results reveal a progressive increase in difficulty from modality-level attribute recognition to higher-level pragmatic attribution, while showing that the attribute heads provide directly readable intermediate semantic predictions for the final sarcasm decision.

C.6 Ablation Studies

We use a progressive ablation to separate the roles of semantic reconstruction and the remaining fine-grained semantic supervision. Every variant uses the same deduplicated joint training and validation source, random seed, checkpoint-selection rule, model architecture, and fixed decision prototypes, and is evaluated on the four complete Golden test sets. We first remove the semantic reconstruction objective to test whether explanation generation accounts for detection performance. Starting from this variant, we then remove the four fine-grained classification losses, label-supervised contrastive learning, and teacher–student KL, leaving only the Strict/Broad detection objectives.

Table C4: Progressive joint-training ablations on the four complete Golden test sets. Each entry reports Strict/Broad Macro-F1 (%); Average is the unweighted mean over the four test sets. All variants use the same deduplicated joint training and validation source, seed, and checkpoint selection rule.

Variant	MMSD	MMSD2.0	SarcNet-EN	SarcNet-ZH	Average
Full SarcAtlas	**82.52/80.86**	**79.97/80.14**	75.19/**80.72**	74.00/76.65	77.92/**79.59**
No semantic reconstruction	82.15/80.57	79.29/79.26	**78.10**/79.65	73.31/**78.74**	**78.21**/79.55
Detection only	80.81/80.32	77.85/78.90	75.70/78.52	**75.98**/77.58	77.59/78.83

The full model and its no-reconstruction variant have similar detection performance: their four-test averages are 77.92 versus 78.21 for Strict and 79.59 versus 79.55 for Broad. The direction of the change also varies across test sets. Semantic reconstruction therefore does not primarily act as an additional detector; instead, it maps the latent semantic states to readable descriptions of shared context, modality-unique information, pragmatic mechanism, and cross-modal relations. The case study in the main paper shows how this mapping makes the intermediate evidence directly inspectable.

After controlling for reconstruction, the comparison with the detection-only variant isolates the aggregate effect of the remaining fine-grained semantic objectives. Fine-grained supervision raises Strict Macro-F1 by 1.34, 1.44, and 2.40 points on MMSD, MMSD2.0, and SarcNet-EN, respectively. It also improves Broad Macro-F1 on all four test sets, by 0.25, 0.36, 1.13, and 1.16 points, yielding average gains of 0.63 Strict points and 0.72 Broad points.

These gains match the roles of the retained objectives. The attribute losses provide discrete anchors for text sarcasm, image sarcasm, evidence source, and pragmatic type; label-supervised contrast organizes the corresponding latent spaces; and teacher–student KL preserves continuous semantic relations from the reconstructed fields. Together, they give the final decision state structured intermediate evidence rather than requiring it to recover every factor from the detection labels alone. The effect is most consistent for Broad detection, whose wider decision boundary directly benefits from explicit modality-level and pragmatic evidence. SarcNet-ZH exhibits a different Strict trade-off: detection-only training is 2.67 points higher on Strict, whereas fine-grained supervision remains 1.16 points higher on Broad. Overall, the progressive ablation separates two complementary roles: fine-grained semantic supervision structures intermediate states for detection, while reconstruction makes those states readable and auditable.

C.7 Comparison Methods and Training Controls

Comparison methods include constant predictors, the mBERT text model, the ViT image model, frozen CLIP fusion, frozen Qwen3-VL representations, and the multimodal sarcasm detectors HFM, Multi-View CLIP, DT4MID (Tomás et al., 2023), and RCLMuFN (Wang et al., 2025). All controlled baselines use the same data partitions, Strict/Broad supervision, random seed, and checkpoint-selection rule. Each trainable model learns only from the training split of its source and selects its best checkpoint by Strict Macro-F1 on the validation split of that same source. After selection, the checkpoint is transferred directly to all four complete test sets.

C.8 Metrics and Evaluation Protocol

Our primary metrics are Strict Macro-F1 and Broad Macro-F1. Strict measures canonical sarcasm with an explicit pragmatic contrast or implicit stance reversal, whereas Broad additionally covers ridicule, satire, exaggeration, and other forms of indirect criticism. All metrics are computed on the complete official test set of each dataset. Cross-dataset experiments use only the validation set of the training source for checkpoint selection: diagonal entries denote in-domain performance, off-diagonal entries denote direct transfer without target-domain adaptation, and the joint-training row reports the model trained on the deduplicated multi-source training set.

Result aggregation and uncertainty estimation. Unless otherwise stated, results aggregated across test sets are equal-weight arithmetic means over the evaluation units displayed in the corresponding table, rather than averages weighted by dataset size. The unified-training and source-specific in-domain evaluations each contain four test units, whereas the cross-dataset transfer evaluation contains eight transfer directions. The 95% confidence intervals reported in the captions of the three main tables are estimated for the average performance difference between SarcAtlas and the specified comparator using 20,000 paired bootstrap replicates. Within each replicate, examples are resampled with replacement within each target test set. When multiple transfer directions share the same target test set, they reuse the same resampling indices to preserve their paired dependence. We take the 2.5th and 97.5th percentiles of the bootstrap distribution as the interval bounds.

For a test set with N negative and P positive examples, an always-negative predictor obtains Macro-F1 $N/(2N + P)$, and an always-positive predictor obtains $P/(N + 2P)$. We include both constant predictors in the main comparison to show the reference level induced by each test distribution.

C.9 Efficiency Measurement

We measure inference efficiency on a fixed set of 256 complete MMSD2.0 test records. All methods use the same sample order, BF16 precision, and one NVIDIA 32 GB NVIDIA GPU. For each batch size, we first run 20 warmup steps and then measure at least 100 steps, with no less than 15 seconds of accumulated forward time. Batch sizes are profiled once in ascending order and once in descending order. Table 4 reports the batch-size-4 results. Latency and throughput cover the end-to-end inference path. During classification inference, SarcAtlas does not execute the frozen reconstruction decoder or any other branch used only during training.