

Structured Semantic Prototype-Guided Multi-Source Multimodal Sentiment Analysis with Fine-Grained Gating and Disagreement Decomposition

Linxuan Song

lxuan0620@163.com

Shangdong Agriculture and Engineering University

Article

Keywords:

Posted Date: October 20th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10693482/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Structured Semantic Prototype-Guided Multi-Source Multimodal Sentiment Analysis with Fine-Grained Gating and Disagreement Decomposition

Linxuan Song^{1*}

^{1*}School of Arts, Shandong Agriculture and Engineering University,
Jinan, 250100, Shandong, China.

Corresponding author(s). E-mail(s): lxuan0620@163.com;

Abstract

Multimodal social-media posts convey sentiment through user text, visual scenes, text embedded in images, and contextual information that may be absent from either primary modality. Existing image-text sentiment models often use coarse modality weighting and unified fusion, limiting their ability to distinguish aligned, complementary, and disagreeing evidence. We propose Structured Semantic Prototype-Guided Multimodal Sentiment Analysis (SPG-MSA), which jointly models user text, visual patches, optical character recognition (OCR) text, and factual image captions. Eight structured evaluative prototypes condition local evidence extraction, an auxiliary-quality-aware gating module selects informative tokens and patches, and three relation-specific branches represent shared, complementary, and disagreement-weighted information. On MVSA-Single and TumEmo, SPG-MSA achieved accuracies of 82.47% and 72.96% and macro-F1 scores of 81.72% and 72.41%, respectively. It exceeded the strongest recent reproduced baseline by 0.84 and 0.85 macro-F1 percentage points. Equal-information controls, prototype validation, source perturbations, duplicate-filtered sensitivity analysis, explanation-fidelity evaluation, and computational profiling showed that the gains were not attributable solely to additional inputs or model capacity. The framework therefore provides a structured approach to multimodal sentiment analysis with improved robustness under controlled perturbations and higher-fidelity evidence attribution.

Keywords: multimodal sentiment analysis, semantic prototype, fine-grained gating, modality disagreement, optical character recognition, multi-source fusion

047 1 Introduction

048 Social-media posts increasingly combine written descriptions, photographs, screen-
049 shots, graphic layouts, packaging information, and text embedded within images.
050 Sentiment in such content is rarely concentrated in a single source. A neutral sentence
051 may accompany a strongly affective image, whereas explicit evaluative language may
052 be paired with visual content that is irrelevant, ambiguous, or affectively inconsistent.
053 Accurate multimodal sentiment analysis therefore requires more than the concate-
054 nation of image and text features. A model must identify which local elements are
055 relevant, determine how evidence from different sources relates, and preserve use-
056 ful modality-specific information without allowing unreliable sources to dominate the
057 prediction.
058

059 The semantic basis of sentiment is also heterogeneous. A positive evaluation may
060 arise from trust, authenticity, consumption experience, nostalgia, production qual-
061 ity, cultural identification, or successful renewal. Research on brand authenticity
062 and heritage-oriented communication illustrates how these evaluative dimensions can
063 jointly shape affective responses [1–3]. Although the present study does not treat brand
064 communication as its empirical domain, these concepts motivate a broader inventory of
065 structured evaluative semantics that can also occur in general social-media discourse.

066 Image-text sentiment analysis has advanced through attention-based fusion,
067 modality calibration, external sentiment-knowledge injection, shared-private represen-
068 tation learning, graph interaction, and fine-grained cross-modal alignment [4–10]. More
069 recent studies have addressed missing modalities, semantic ambiguity, and robustness
070 under incomplete or inconsistent inputs [11–14].

071 A further line of research has introduced explicit semantic completion and more
072 structured fusion. Visual-elements-as-prompts methods have also shown that in-image
073 text and generated visual descriptions can support multimodal sentiment recognition
074 [15]. CoDe uses text embedded in images to complete image and textual semantics
075 and decomposes modality-shared and modality-private sentiment representations [16].
076 DualScope combines global-local feature fusion with bidirectional image-text genera-
077 tion and contrastive learning [17]. SSEDf dynamically integrates shared and private
078 semantic representations [18], whereas CLIP-MoE adapts a contrastively pretrained
079 vision-language model through mixture-of-experts modules, supervised cross-modal
080 alignment, and online logit modulation [19]. The conceptual distinctions among these
081 approaches and SPG-MSA are summarized in Supplementary Table S1.

082 Despite these advances, three limitations remain. First, most methods are super-
083 vised primarily by coarse sentiment or emotion categories. Such labels identify an
084 affective outcome but do not describe the semantic basis on which the judgment is
085 formed. Second, adaptive fusion is frequently performed using one pooled representa-
086 tion or one coefficient for an entire modality. A global coefficient cannot distinguish
087 an informative adjective from an unrelated hashtag, or a sentiment-bearing sign from
088 a visually dominant but irrelevant background. Fine-grained studies have already
089 shown the value of text-guided interaction and local alignment [20–22]. Third, cross-
090 modal variation is often reduced to shared and private information or treated as

091
092

noise. Modality-specific evidence may, however, be complementary without being contradictory, while disagreement between preliminary unimodal predictions may signal irrelevant, uncertain, or affectively inconsistent evidence.

To address these limitations, we propose Structured Semantic Prototype-Guided Multimodal Sentiment Analysis (SPG-MSA). The framework retains the original image-text observations and sentiment labels while deriving two auxiliary sources from each image: optical-character-recognition (OCR) text and a factual image caption. Structured evaluative prototypes guide local evidence extraction. Reliability indicators regulate the automatically generated auxiliary sources, and scalar gates select relevant tokens and image patches. The primary image and text representations are subsequently organized into shared, complementary, and disagreement-weighted branches.

The main contributions are threefold. First, we introduce a structured evaluative semantic-prototype mechanism that conditions multimodal evidence extraction through interpretable concepts rather than output-class prototypes. Second, we develop an auxiliary-quality-aware fine-grained gating mechanism that combines content relevance with OCR confidence and image-caption relevance. Third, we propose a relation-specific representation framework that separately models shared, complementary, and disagreement-weighted information, and we evaluate it using controlled comparison, ablation, source perturbation, calibration, explanation fidelity, and efficiency analyses.

2 Methods

2.1 Task definition and semantic-prototype inventory

Let the multimodal dataset be

$$\mathcal{D} = \{(T_i, I_i, y_i)\}_{i=1}^N, \quad (1)$$

where T_i is the user-written text, I_i is the corresponding image, $y_i \in \{1, \dots, K\}$ is the sentiment or emotion label, N is the number of observations, and K is the number of classes. Equation 1 defines the supervised observations used throughout the study.

Two auxiliary sources are derived from the original image:

$$O_i = \text{OCR}(I_i), \quad C_i = \text{Caption}(I_i), \quad (2)$$

where O_i denotes OCR-extracted text and C_i is a factual description of visible image content. The prediction function is

$$\hat{y}_i = f_{\Theta}(T_i, I_i, O_i, C_i, \mathcal{A}), \quad (3)$$

where Θ denotes trainable parameters and $\mathcal{A} = \{a_1, \dots, a_A\}$ is the semantic-prototype inventory with $A = 8$. Equations 2 and 3 make clear that the auxiliary sources are derived from the original image rather than additional labelled observations.

The eight structured evaluative prototypes cover temporal continuity, cultural identity, production and craftsmanship, perceived authenticity, nostalgia and identification, trust and reputation, contemporary renewal, and consumption experience. Their operational definitions and representative evidence are given in Supplementary Table S2. The prototypes act as latent priors; the original sentiment or emotion category remains the only sample-level supervised target.

2.2 Framework overview

Figure 1 shows the complete SPG-MSA architecture. The model contains five stages: multi-source extraction and encoding, structured semantic-prototype conditioning, auxiliary-quality-aware local gating, relation-specific representation, and prototype-conditioned hierarchical fusion and classification.

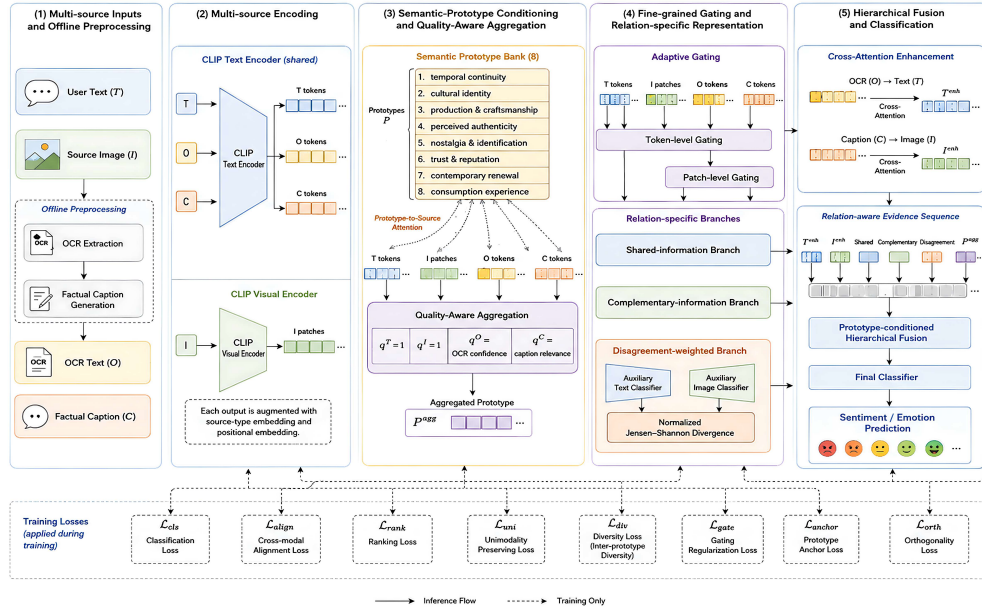


Fig. 1 Overall architecture of SPG-MSA. Four sources are encoded in a shared visual-language space, conditioned by structured semantic prototypes, locally gated, decomposed into relation-specific representations, and fused for sentiment or emotion prediction. Solid connections indicate inference-time operations, whereas dashed connections denote training-only objectives. OCR extraction and factual caption generation are performed offline.

2.3 Auxiliary-source extraction

2.3.1 OCR extraction

OCR is performed using PP-OCRv4 [23]. Text boxes with confidence below 0.65 are discarded. The length-weighted OCR confidence is

$$q_i^O = \begin{cases} \frac{\sum_{b=1}^{B_i} l_{i,b} c_{i,b}}{\sum_{b=1}^{B_i} l_{i,b}}, & B_i > 0, \\ 0, & B_i = 0, \end{cases} \quad (4)$$

where B_i is the number of retained text boxes, $l_{i,b}$ is the recognized-string length, and $c_{i,b}$ is recognition confidence. Equation 4 assigns zero quality when no valid OCR box is available; such samples use a special [NO_OCR] token.

2.3.2 Factual image captioning

A frozen BLIP-2 OPT-2.7B model [24] generates one caption for each image using greedy decoding and at most 30 new tokens. The prompt is: ‘‘Describe the visible entities, actions, objects, scene, and readable text in the image. Do not infer emotions, sentiment, intentions, or opinions.’’ All captions are generated before model training and cached.

Image-caption relevance is measured as

$$q_i^C = \frac{1 + \cos(\hat{x}_i^I, \hat{x}_i^C)}{2}, \quad (5)$$

where $\hat{x}_i^I$ and $\hat{x}_i^C$ are L2-normalized pooled CLIP representations. Equation 5 measures semantic correspondence rather than complete factual accuracy. User text and the original image are treated as primary observations and assigned unit priors:

$$q_i^T = q_i^I = 1. \quad (6)$$

Equation 6 distinguishes the primary observations from the automatically generated auxiliary sources.

2.4 Multi-source encoding

A pretrained Contrastive Language-Image Pretraining (CLIP) ViT-B/16 model [25] provides the shared visual-linguistic representation space. For each source $m \in \{T, I, O, C\}$,

$$X_i^m = E_m(S_i^m) \in \mathbb{R}^{L_m \times d}, \quad (7)$$

where S_i^m is the source input, L_m is its sequence length, and $d = 512$. Equation 7 yields token sequences for text-based sources and patch sequences for the image.

The encoded sequence is projected and augmented as

$$\bar{X}_i^m = \text{LN}(X_i^m W_m + b_m + E_{\text{src}}^m + E_{\text{pos}}^m), \quad (8)$$

where E_{src}^m is a learned source-type embedding and E_{pos}^m is a positional embedding. Equation 8 preserves source identity while placing all representations in the same fusion space; padded positions are masked.

234

235 2.5 Prototype construction and semantic conditioning

For each semantic prototype a_j , four prompt templates are encoded. The initial prototype is

$$239 \quad p_j^0 = \frac{1}{R} \sum_{r=1}^R E_T(q_{j,r}), \quad R = 4, \quad (9)$$

where $q_{j,r}$ is the r th prompt for prototype j . Equation 9 initializes the trainable prototype matrix $P = [p_1; \dots; p_A] \in \mathbb{R}^{A \times d}$ from pretrained language semantics.

Prototype-to-source attention is

$$245 \quad S_i^m = \text{softmax} \left[\frac{PW_Q(\bar{X}_i^m W_K)^\top}{\sqrt{d_h}} + M_i^m \right], \quad H_i^m = S_i^m \bar{X}_i^m W_V, \quad (10)$$

where M_i^m masks padding and d_h is the attention-head dimension. Equation 10 produces source-specific prototype-conditioned features.

Prototype drift is controlled by

$$253 \quad \mathcal{L}_{\text{anchor}} = \frac{1}{A} \sum_{j=1}^A [1 - \cos(p_j, p_j^0)], \quad (11)$$

and collapse is discouraged by

$$258 \quad \mathcal{L}_{\text{div}} = \left\| \hat{P} \hat{P}^\top - I_A \right\|_F^2, \quad (12)$$

where $\hat{P}$ is row-normalized. Equations 11 and 12 respectively preserve semantic anchoring and encourage inter-prototype diversity.

263

264 2.6 Auxiliary-quality-aware prototype aggregation

Let $\bar{h}_i^m = \text{MeanPool}(H_i^m)$. The contribution of source m is

$$268 \quad \beta_i^m = \frac{\exp(u^\top \bar{h}_i^m + \delta_m q_i^m)}{\sum_n \exp(u^\top \bar{h}_i^n + \delta_n q_i^n)}, \quad H_i^A = \sum_m \beta_i^m H_i^m. \quad (13)$$

Equation 13 combines learned content relevance with observable OCR confidence or image-caption relevance. User text and the source image retain the unit priors defined in Equation 6.

274

275

276

2.7 Fine-grained adaptive gating

For each source, contextual evidence from the remaining sources is obtained through

$$R_i^m = \text{MHA}(Q = \bar{X}_i^m, K = \bar{X}_i^{-m}, V = \bar{X}_i^{-m}), \quad (14)$$

where $\bar{X}_i^{-m}$ denotes the masked concatenation of the other sources. Equation 14 allows each local token or patch to be evaluated against cross-source context.

The semantic-prototype context is repeated across the source sequence and combined with $\bar{X}_i^m$ and R_i^m . A scalar gate is then assigned to every textual token or visual patch:

$$G_i^m = \sigma \left(\frac{W_g^m [\bar{X}_i^m \| R_i^m \| B_i^m] + b_g^m}{\tau} \right), \quad G_i^m \in \mathbb{R}^{L_m \times 1}, \quad (15)$$

where B_i^m is the repeated pooled prototype context and τ is the gate temperature. The gated representation is

$$Z_i^m = G_i^m \odot \bar{X}_i^m + (1 - G_i^m) \odot r_m, \quad (16)$$

where r_m is a learned residual embedding. Equations 15 and 16 implement token- and patch-level evidence selection.

Gate collapse is discouraged by

$$\mathcal{L}_{\text{gate}} = \sum_m \left[\frac{1}{BL_m} \sum_{i=1}^B \sum_{l=1}^{L_m} G_{i,l}^m - \rho_m \right]^2, \quad (17)$$

where B is batch size and ρ_m is the target activation ratio. Equation 17 keeps average gate activity close to a validation-selected target.

2.8 Relation-specific representation

Figure 2 summarizes the shared, complementary, and disagreement-weighted branches. The branch names describe optimization roles rather than mathematically unique latent factors.

Bidirectional text-image attention yields a shared representation

$$h_i^{\text{sh}} = \frac{\text{Pool}[\text{MHA}(Z_i^T, Z_i^I, Z_i^I)] + \text{Pool}[\text{MHA}(Z_i^I, Z_i^T, Z_i^T)]}{2}. \quad (18)$$

Equation 18 captures information that can be aligned across the primary text and image sources.

Let $z_i^T = \text{Pool}(Z_i^T)$ and $z_i^I = \text{Pool}(Z_i^I)$. The complementary representation is

$$h_i^{\text{cp}} = \text{MLP} \left[(z_i^T - W_{TS} h_i^{\text{sh}}) \| (z_i^I - W_{IS} h_i^{\text{sh}}) \right]. \quad (19)$$

Equation 19 retains modality-specific residual information after subtracting projected shared content.

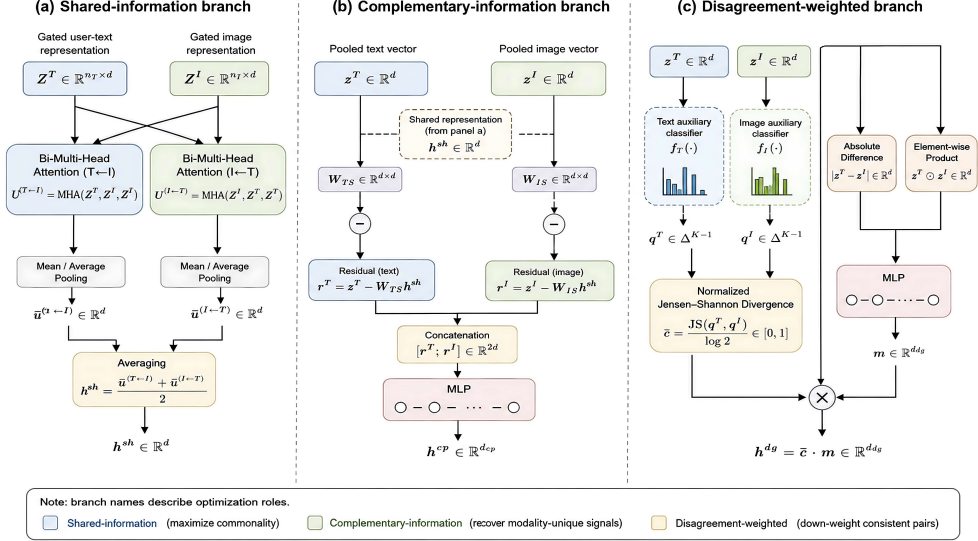


Fig. 2 Relation-specific representation in SPG-MSA. The shared branch uses bidirectional cross-attention, the complementary branch removes projected shared information to retain modality-specific residuals, and the disagreement branch weights interaction features by normalized Jensen-Shannon divergence between preliminary unimodal predictions.

Two auxiliary classifiers produce preliminary distributions q_i^T and q_i^I . With $m_i = (q_i^T + q_i^I)/2$, normalized Jensen-Shannon divergence [26] is

$$\bar{c}_i = \frac{\frac{1}{2}D_{\text{KL}}(q_i^T \| m_i) + \frac{1}{2}D_{\text{KL}}(q_i^I \| m_i)}{\ln 2}, \quad 0 \leq \bar{c}_i \leq 1. \quad (20)$$

Equation 20 quantifies disagreement between preliminary sentiment distributions rather than general semantic incompatibility. The disagreement-weighted representation is

$$h_i^{\text{dg}} = \bar{c}_i \text{MLP}(|z_i^T - z_i^I| \| z_i^T \odot z_i^I). \quad (21)$$

Equation 21 increases the contribution of interaction features when unimodal predictions diverge.

2.9 Multi-source hierarchical fusion

OCR evidence supplements user text through text-to-OCR cross-attention, and the factual caption supplements the image through image-to-caption cross-attention. The resulting text-enhanced and visual-enhanced vectors are concatenated with h_i^{sh} , h_i^{cp} , and h_i^{dg} to form a relation-aware evidence sequence F_i . Prototype-conditioned fusion and classification are

$$\tilde{h}_i = \text{MHA}[\text{Pool}(H_i^A), F_i, F_i],$$

$$\begin{aligned} h_i^* &= \text{LN}[\tilde{h}_i + W_f \text{Pool}(F_i)], \\ \hat{y}_i &= \text{softmax}(W_c h_i^* + b_c). \end{aligned} \quad (22)$$

Equation 22 produces the final sentiment or emotion distribution from prototype-conditioned multi-source evidence.

2.10 Training objectives

Class imbalance is handled using class-weighted cross-entropy:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N \frac{N}{KN_{y_i}} \log \hat{y}_{i,y_i}, \quad (23)$$

where N_{y_i} is the number of training observations in the class of sample i . Equation 23 avoids duplicating observations through oversampling.

Instance-level image-text alignment is bidirectional:

$$\mathcal{L}_{\text{ins}} = \frac{1}{2}(\mathcal{L}_{T \rightarrow I} + \mathcal{L}_{I \rightarrow T}), \quad (24)$$

where each directional contrastive loss treats the original image-text pair as positive. Equation 24 is combined with a supervised label-level contrastive objective following the general principle of supervised contrastive learning [27]. Let $\mathcal{P}(i) = \{p \neq i : y_p = y_i\}$ and $\mathcal{I}_+ = \{i : |\mathcal{P}(i)| > 0\}$. Then

$$\mathcal{L}_{\text{lab}} = -\frac{1}{|\mathcal{I}_+|} \sum_{i \in \mathcal{I}_+} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{\exp[\text{sim}(h_i^*, h_p^*)/\gamma]}{\sum_{a \neq i} \exp[\text{sim}(h_i^*, h_a^*)/\gamma]}. \quad (25)$$

Equation 25 excludes anchors without a same-class positive in the mini-batch. A class-aware sampler includes at least two instances per represented class whenever batch size permits. The complete alignment term is $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{ins}} + \omega \mathcal{L}_{\text{lab}}$.

For the three relation-specific branches, row-normalized batch matrices H^a are centered as $\tilde{H}^a = H^a - B^{-1} \mathbf{1} \mathbf{1}^\top H^a$. The cross-covariance regularizer is

$$\mathcal{L}_{\text{orth}} = \sum_{a < b} \left\| \frac{(\tilde{H}^a)^\top \tilde{H}^b}{B-1} \right\|_F^2, \quad (26)$$

where $a, b \in \{\text{sh}, \text{cp}, \text{dg}\}$. Equation 26 discourages redundant branch representations without claiming unique latent-factor identification.

The unimodal classifiers are trained using $\mathcal{L}_{\text{uni}} = \text{CE}(q^T, y) + \text{CE}(q^I, y)$. For a label-incongruent image pair, disagreement ranking is

$$\mathcal{L}_{\text{rank}} = \frac{1}{B} \sum_i \max(0, \mu + \bar{c}_i^+ - \bar{c}_i^-), \quad (27)$$

where $\bar{c}_i^+$ and $\bar{c}_i^-$ are disagreement scores for the original and label-incongruent pairs. Equation 27 encourages higher disagreement for label-incongruent pairs.

The complete training objective is

$$\begin{aligned} \mathcal{L} = & \mathcal{L}_{\text{cls}} + \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{orth}} + \lambda_3 \mathcal{L}_{\text{uni}} + \lambda_4 \mathcal{L}_{\text{rank}} \\ & + \lambda_5 \mathcal{L}_{\text{gate}} + \lambda_6 \mathcal{L}_{\text{anchor}} + \lambda_7 \mathcal{L}_{\text{div}}. \end{aligned} \quad (28)$$

Equation 28 jointly optimizes classification, alignment, branch separation, unimodal supervision, disagreement ranking, gate activity, and prototype regularization.

2.11 Training with unavailable sources

Source dropout is applied during training. When a textual source is dropped, its sequence is replaced by a learned [MISSING_SOURCE] token and its quality or relevance score is set to zero. When the image is dropped, the visual patch sequence is replaced by a learned missing-image representation. Dropped sources are excluded from cross-attention keys and values through attention masks. This procedure improves robustness to incomplete input but does not create additional labelled observations.

2.12 Datasets and fixed study splits

MVSA-Single contains 5,129 Twitter image-text pairs whose image and text modalities were independently annotated as positive, neutral, or negative [28]. Following the commonly used preprocessing protocol, samples with matching modality labels were retained; when one modality was neutral and the other was positive or negative, the non-neutral label was used; directly opposing positive-negative pairs were removed. The processed dataset contained 4,511 observations.

TumEmo is a weakly supervised image-text emotion dataset collected from Tumblr and contains seven classes: angry, bored, calm, fear, happy, love, and sad [29]. All splits were generated once before training using fixed stratified identifiers. The same identifiers were used for every reproduced method, and no model-selection or hyperparameter decision used the test partitions. The complete class distribution is reported in Supplementary Table S3.

2.13 Duplicate audit and sensitivity analysis

Exact textual duplicates were detected after lowercasing, URL normalization, user-mention replacement, and whitespace normalization. Near-duplicate text was identified using Sentence-BERT embeddings [30] with cosine similarity above 0.95. Exact image duplicates were identified using cryptographic hashes; near-duplicate images were detected using perceptual hashing [31] with a Hamming-distance threshold of five. Duplicate categories were not mutually exclusive. Detailed counts are reported in Supplementary Table S10.

The fixed study split was retained for the primary evaluation. A duplicate-filtered sensitivity analysis removed affected test observations without retraining the model. This analysis assesses whether duplicate test content inflated the reported scores; it

not equivalent to reconstructing a fully group-aware split. The sensitivity results are reported in Supplementary Table S11.

2.14 Baselines and equal-information controls

The conventional comparison methods were Image-CLIP, Text-CLIP, Early Fusion, MVCN [5], SKEAFN-IT [6], SPMN-IT [7], CIGNN [9], MMJL [10], and MFN [4]. Four recent methods were additionally reproduced: CoDe-R [16], DualScope-R [17], SSEDf-R [18], and CLIP-MoE-R [19]. The suffix “-R” denotes controlled reproduction. When a method had not originally been evaluated on TumEmo, its output head and label dimensions were adapted without altering the principal fusion mechanism. DualScope-R was not transferred to TumEmo because its bidirectional generation and augmentation pipeline was designed for the three-class MVSA setting.

All comparison methods used identical split identifiers and evaluation scripts, and each received the same validation-based hyperparameter-search budget of 20 configurations. Four equal-information controls received the same four sources as SPG-MSA: concatenation, a four-source Transformer, four-source cross-attention, and a parameter-matched four-source Transformer. The parameter-matched model remained within 1.5% of the trainable-parameter count of SPG-MSA.

2.15 Implementation details

CLIP ViT-B/16 served as the backbone. User text was truncated to 77 tokens, OCR text and captions to 40 tokens, and images were resized to 224×224 . The model used eight attention heads, two fusion layers, a feed-forward dimension of 1,024, and dropout of 0.20. The gate temperature was 0.07, contrastive temperature 0.10, and ranking margin 0.20.

Optimization used AdamW [32]. The pretrained backbone learning rate was 1×10^{-5} , the learning rate for new modules was 1×10^{-4} , and weight decay was 1×10^{-4} . Batch sizes were 32 for MVSA-Single and 64 for TumEmo. Training continued for at most 30 epochs with early stopping after five validation epochs without macro-F1 improvement. Gradient norms were clipped at 1.0. The loss coefficients were $(\lambda_1, \dots, \lambda_7) = (0.10, 0.05, 0.10, 0.10, 0.01, 0.02, 0.01)$. Gate targets were $\rho_T = 0.45$, $\rho_I = 0.35$, $\rho_O = 0.30$, and $\rho_C = 0.35$. OCR and caption dropout probabilities were 0.10; text and image dropout probabilities were 0.05. Experiments used an NVIDIA RTX 3090 and random seeds {13, 21, 42, 67, 89}. Unless stated otherwise, results are means $\pm$ standard deviations across five independent runs. Parameter sensitivity is provided in Supplementary Table S14.

2.16 Controlled source perturbations

Three image-mismatch conditions were constructed. For random cross-pair mismatch, the image was replaced by one from another post after excluding the original pair and detected duplicate groups. For same-label semantic mismatch, the replacement image was drawn from the same class but from the lowest quartile of CLIP similarity to the original text. For cross-valence semantic-neighbor mismatch in MVSA-Single, the replacement was drawn from the opposite positive or negative class and chosen as the

candidate with the highest CLIP similarity; neutral observations were excluded. For TumEmo, happy and love formed the positive group, angry, fear, and sad the negative group, and calm and bored were excluded from cross-valence sampling. Perturbation lists were generated once and shared by all five seeds. OCR corruption randomly replaced retained OCR tokens, whereas sentiment-word masking used the NRC Emotion Lexicon [33]. Accuracy across corruption ratios is reported in Supplementary Table S13.

2.17 Human evaluation, caption audit, and statistical analysis

Two hundred test samples from each dataset were independently evaluated by three annotators with experience in sentiment analysis. Annotators selected one or more applicable semantic prototypes or “none of the above or insufficient evidence” and were blinded to model predictions. Multi-label Krippendorff’s alpha [34] was computed separately for each prototype and macro-averaged; majority-vote annotations served as reference labels.

For caption auditing, 1,000 captions from each dataset were assessed for factual correctness, hallucinated objects or actions, and omission of visually central objects. Explicit affective words were detected using the NRC Emotion Lexicon [33]. The sentiment-oriented caption prompt was used only as a leakage-control condition.

Accuracy and macro-F1 were the primary classification metrics. Calibration was evaluated using expected calibration error (ECE) with ten equal-width bins [35] and raw softmax probabilities. Explanation fidelity used comprehensiveness and sufficiency following perturbation-based rationale evaluation [36]. For correctly classified observations, the top 20% of user-text tokens and image patches were selected; removed text tokens were replaced by padding tokens and removed visual patches by the dataset-level mean patch. Random evidence selection was repeated 20 times per sample. Prototype geometry was visualized using UMAP [37].

Macro-F1 differences were evaluated with 10,000-iteration hierarchical stratified bootstrap resampling [38]. Each iteration sampled a random seed and then resampled test observations within class. Holm correction [39] was applied across baseline comparisons, with adjusted $p < 0.05$ considered statistically reliable.

3 Results

3.1 Overall performance

The complete controlled comparison is reported in Supplementary Table S4. SPG-MSA achieved accuracies of 0.8247 and 0.7296 and macro-F1 scores of 0.8172 and 0.7241 on MVSA-Single and TumEmo, respectively. Relative to MFN, macro-F1 increased by 5.04 percentage points on MVSA-Single and 4.97 points on TumEmo.

The comparison with recent methods was more competitive. Relative to CLIP-MoE-R, SPG-MSA improved macro-F1 by 0.84 percentage points on MVSA-Single and 0.85 points on TumEmo, corresponding to relative macro-F1 error reductions of 4.39% and 2.99%. Hierarchical-bootstrap differences were 0.0084 on MVSA-Single (95% CI [0.0021, 0.0147], Holm-adjusted $p = 0.014$) and 0.0085 on TumEmo (95%

CI [0.0042, 0.0128], adjusted $p = 0.006$). The absolute advantage over CLIP-MoE-R remained below one percentage point and is therefore interpreted as statistically reliable but moderate.

SPG-MSA exceeded CoDe-R by 1.54 and 1.55 macro-F1 points. Because both methods use OCR evidence and representation decomposition, this comparison suggests that structured semantic conditioning, factual-caption evidence, auxiliary-source regulation, and local gating provide benefits beyond in-image text completion and shared-private decomposition alone.

3.2 Equal-information controls and ablation

Supplementary Table S5 separates improvements arising from additional source information, parameter count, and individual architectural components. Simple concatenation of all four sources yielded macro-F1 scores of 0.7769 and 0.6830, showing that OCR and captions did not automatically improve prediction. The parameter-matched Transformer remained 1.48 and 1.55 points below SPG-MSA on MVSA-Single and TumEmo, respectively.

Fine-grained gating produced the largest individual ablation effect. Replacing token- and patch-level gates with a single source coefficient reduced macro-F1 by 2.65 and 2.47 points. Removing the disagreement-weighted branch reduced macro-F1 by 1.55 and 1.59 points. These results indicate that local evidence selection and explicit representation of affective divergence contribute beyond model capacity and source count.

3.3 Semantic-prototype validation

Computational prototype controls and human relevance evaluation are reported in Supplementary Table S6. Random frozen vectors produced only marginal gains over the model without prototypes. Unnamed learnable vectors performed better but remained 1.36 and 1.40 points below the complete structured prototypes. Sentiment-label and shuffled prototypes also underperformed the complete configuration.

Generic brand-attribute prototypes were more competitive and remained only 0.70 and 0.68 points below the complete model. This comparatively small margin indicates that part of the gain arises from structured evaluative semantics in general rather than exclusively from the selected heritage-inspired inventory.

Human evaluation produced Krippendorff’s alpha values of 0.781 and 0.742. The highest-ranked model prototype matched the majority annotation for 71.4% of MVSA-Single observations and 68.1% of TumEmo observations; Hit@3 exceeded 0.85 on both datasets. The lower TumEmo result is consistent with its emotion-centered construction, in which some posts contain facial or situational affect without a clear temporal, cultural, reputational, or consumption-related dimension.

3.4 Auxiliary-source contribution and caption audit

Supplementary Table S7 reports source contributions, OCR availability, and caption quality. OCR was retained in 42.6% of MVSA-Single images and 29.4% of TumEmo images. It therefore functioned as a sparse auxiliary source rather than a consistently

available modality. Factual captions improved both datasets by translating scene-level visual information into a language-compatible representation.

Removing auxiliary-quality signals reduced macro-F1 by 0.83 points on MVSA-Single and 1.03 points on TumEmo. The larger reduction on TumEmo is consistent with its lower mean image-caption similarity and higher hallucination rate.

Figure 3 summarizes the caption analysis. The sentiment-oriented prompt produced stronger caption-only performance, showing that a frozen caption model can introduce external affective inference. Removing explicit affective words from factual captions, however, reduced complete-model macro-F1 by only 0.13 points on each dataset, indicating that most of the caption contribution arose from scene, action, and object semantics rather than direct sentiment-label leakage.

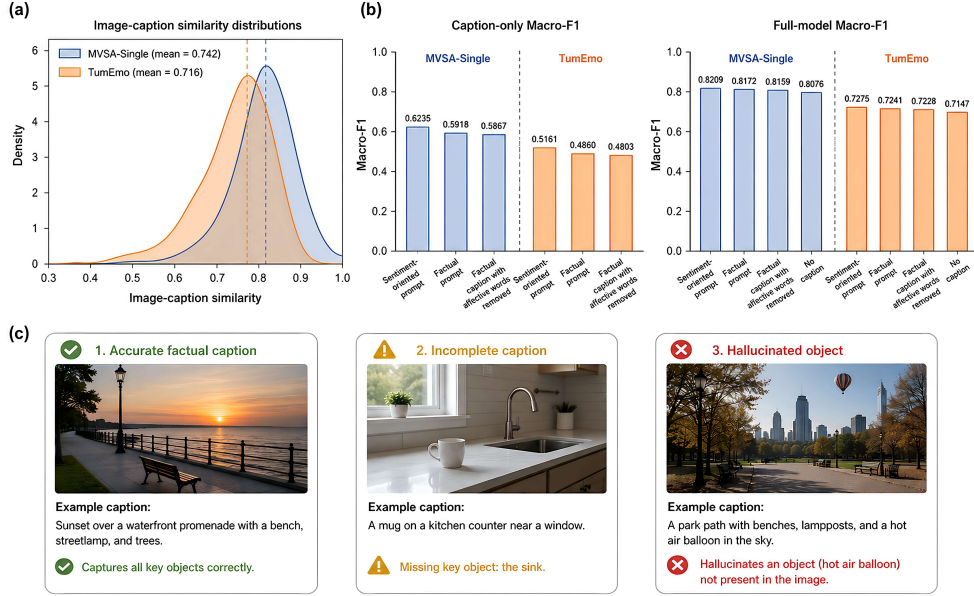


Fig. 3 Factual-caption analysis. (a) Image-caption similarity distributions for MVSA-Single and TumEmo. (b) Caption-only and full-model macro-F1 under sentiment-oriented, factual, affective-word-removed factual, and no-caption conditions. (c) Representative caption cases illustrating a factually correct caption, an incomplete caption, and a hallucinated object.

3.5 Class-level performance and calibration

Complete class-level precision, recall, F1, standard deviation, and support are reported in Supplementary Table S12. On MVSA-Single, positive, neutral, and negative F1 scores were 0.872 ± 0.005 , 0.765 ± 0.009 , and 0.815 ± 0.007 , respectively. Neutral remained the most difficult class. On TumEmo, happy and love achieved 0.793 ± 0.003 and 0.781 ± 0.003 , whereas fear and bored were the weakest classes at 0.631 ± 0.006 and 0.648 ± 0.006 .

Figure 4 shows the mean row-normalized confusion matrices. The matrices average percentages across five independently trained models rather than using ensemble predictions.

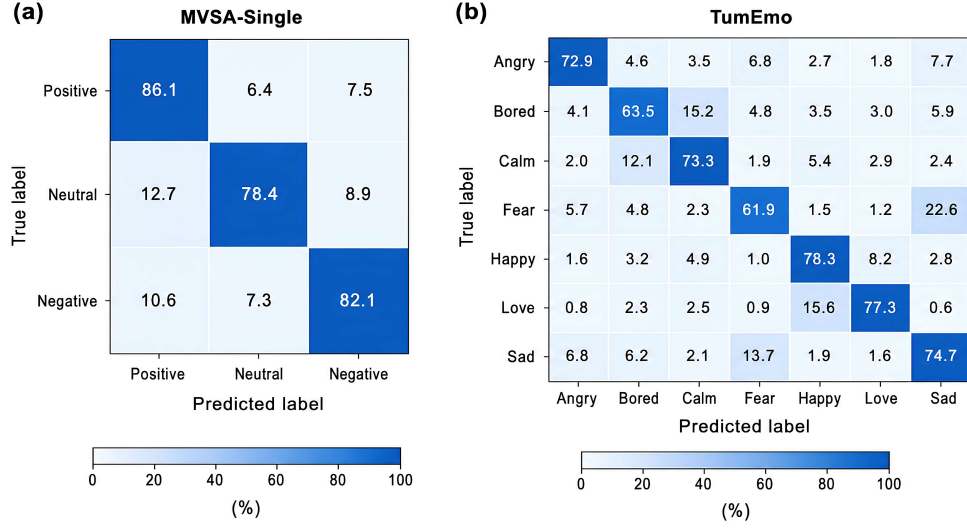


Fig. 4 Mean row-normalized confusion matrices. (a) MVSA-Single positive, neutral, and negative classes. (b) TumEmo angry, bored, calm, fear, happy, love, and sad classes. Each cell reports the mean row-normalized percentage across five independently trained models.

Calibration results are shown in Figure 5. SPG-MSA achieved ECE values of 0.041 and 0.058 on MVSA-Single and TumEmo, compared with 0.067 and 0.081 for MFN. The largest difference occurred in the 0.70–0.90 confidence interval, where MFN exhibited greater overconfidence.

3.6 Missing-source and controlled-mismatch robustness

Robustness under source removal and controlled perturbation is reported in Supplementary Table S8. Removing user text caused the largest decline, especially on TumEmo, where accuracy decreased by 15.70 points. OCR and caption removal had smaller effects because these sources were auxiliary and intermittently available.

Same-label semantic mismatch was less damaging than random cross-pair mismatch because semantically unrelated images could still produce similar preliminary sentiment distributions. Cross-valence semantic-neighbor mismatch was more difficult: accuracy decreased by 6.28 points on MVSA-Single and 5.42 points on TumEmo. Under this condition, removing the disagreement branch further reduced accuracy to 0.7326 and 0.6489, respectively. Thus, disagreement modeling mitigated but did not eliminate semantically plausible affective inconsistency.

Figure 6 shows degradation across increasing corruption ratios, with the underlying values reported in Supplementary Table S13. At a 50% image-mismatch ratio,

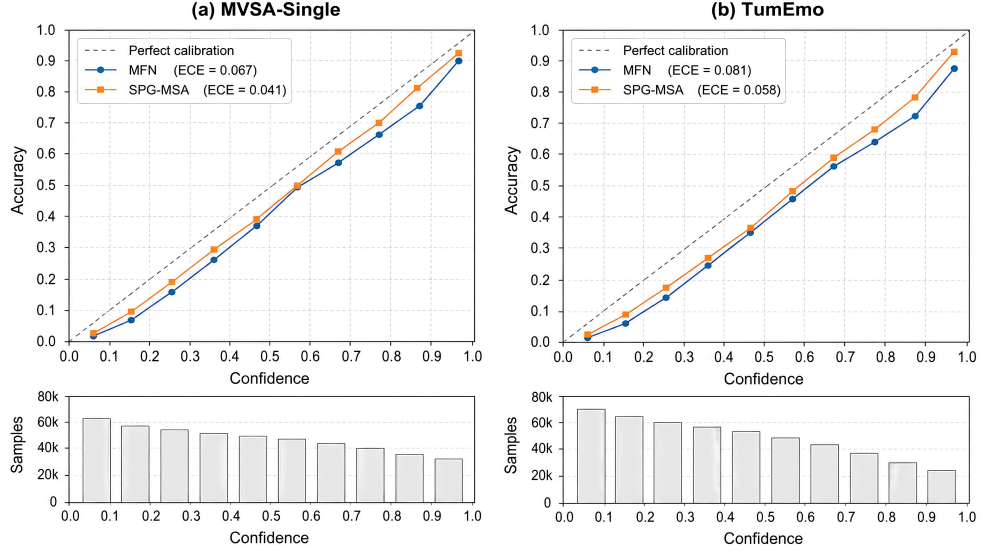


Fig. 5 Reliability diagrams for MFN and SPG-MSA. (a) MVSA-Single and (b) TumEmo results using ten equal-width confidence bins and raw softmax probabilities. The dashed diagonal denotes perfect calibration; background bars show bin sample counts.

SPG-MSA retained accuracies of 0.7210 and 0.6285, whereas the variant without disagreement modeling reached 0.6760 and 0.5821. OCR corruption produced the shallowest degradation, while sentiment-word masking was particularly damaging on TumEmo.

3.7 Prototype geometry

The average off-diagonal cosine similarity among the structured prototypes was 0.18, with the strongest relationship between perceived authenticity and trust and reputation. This proximity is conceptually plausible but also indicates that the inventory is not perfectly separable.

Figure 7 visualizes the prototype similarity matrix and UMAP representation geometry. The UMAP plot is used only as a qualitative visualization and is not treated as evidence of statistically distinct clusters.

3.8 Explanation fidelity and computational efficiency

Perturbation-based explanation fidelity and computational cost are summarized in Supplementary Table S9. Raw cross-attention outperformed random evidence selection but remained weaker than gradient attribution and fine-grained gates. Combining gate importance with prototype relevance produced the strongest comprehensiveness and sufficiency results. Sufficiency nevertheless remained above zero for all methods, indicating that predictions depended on distributed evidence rather than a single decisive region or token.

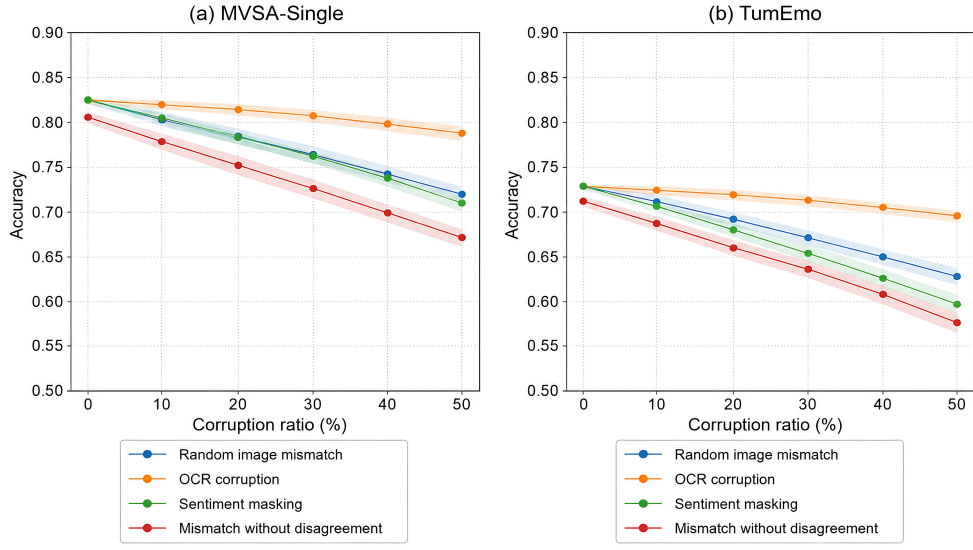


Fig. 6 Robustness under increasing corruption. (a) MVSA-Single and (b) TumEmo accuracy across corruption ratios from 0% to 50% for random image mismatch, OCR corruption, sentiment-word masking, and image mismatch without the disagreement branch. Curves show five-seed means with standard-deviation bands.

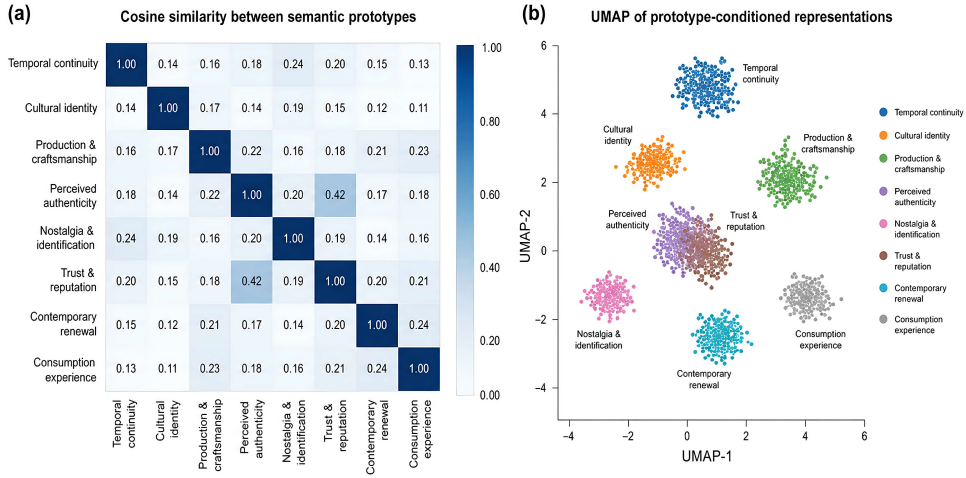


Fig. 7 Prototype geometry. (a) Cosine-similarity matrix for the eight structured evaluative prototypes. (b) UMAP visualization of prototype-conditioned representations. The visualization is descriptive and is not interpreted as proof of statistically distinct clusters.

SPG-MSA used 13.6 million more trainable parameters than MFN, an 8.6% increase. FLOPs increased by 26.9%, peak memory by 24.1%, and training time per epoch by 30.7%. Online latency increased from 12.8 to 18.7 ms per observation. When OCR and caption generation were not cached, end-to-end processing required approximately 1.17 s per sample. The classification and interpretation gains therefore involved a measurable efficiency trade-off.

3.9 Qualitative analysis

Figure 8 presents representative successes and failures. The examples include a product image containing packaging text, weakly affective text accompanied by an informative visual scene, an affectively inconsistent image-text pair, a stylized-OCR error, a caption hallucination, and an irony-related failure. User-identifying information is absent from the displayed cases.

ID	Image (attention overlay)	User Text (snippet)	OCR Output (snippet)	Factual Caption (snippet)	True Label	Pred. Label	Top Prototype Importance (top 5)	Branch Contribution
(a)		Love this soap! Leaves my skin so soft.	NATURAL SOAP / ALOE & OAT / HANDMADE / COLD PROCESS / 120g e 4.2 oz	A bar of natural soap and its box on a wooden surface.	Positive	Positive	- production & craftsmanship: 0.34 - consumption experience: 0.27 - perceived authenticity: 0.16 - trust & reputation: 0.12 - cultural identity: 0.11	Shared (dominant)
(b)		Not feeling great today. Just tired.	—	A calm mountain lake surrounded by trees and peaks.	Positive	Positive	- consumption experience: 0.31 - contemporary renewal: 0.24 - cultural identity: 0.18 - trust & reputation: 0.14 - temporal continuity: 0.13	Complementary (dominant)
(c)		Best day ever! Couldn't be happier.	—	Rainy street viewed through a window with water drops.	Negative	Positive	- temporal continuity: 0.28 - nostalgia & identification: 0.22 - perceived authenticity: 0.18 - cultural identity: 0.16 - contemporary renewal: 0.16	Disagreement (dominant)
(d)		New mug motivation!	Fuel / your / Dreams	A black mug with decorative text on a desk.	Positive	Negative	- perceived authenticity: 0.26 - trust & reputation: 0.21 - consumption experience: 0.19 - production & craftsmanship: 0.18 - nostalgia & identification: 0.16	Disagreement (dominant)
(e)		Simple dinner, so good.	—	A plate of pasta with tomato sauce, herbs, and a glass of wine on a table.	Positive	Negative	- perceived authenticity: 0.25 - consumption experience: 0.23 - production & craftsmanship: 0.18 - temporal continuity: 0.18 - cultural identity: 0.16	Disagreement (dominant)
(f)		I absolutely love Mondays!	I ♥ Mondays	A coffee mug that says "I ♥ Mondays" on an office desk.	Negative	Positive	- nostalgia & identification: 0.27 - temporal continuity: 0.22 - perceived authenticity: 0.18 - cultural identity: 0.17 - trust & reputation: 0.16	Disagreement (dominant)

Fig. 8 Qualitative analysis of SPG-MSA. Six representative cases show anonymized source inputs, predictions, prototype distributions, local evidence maps, and dominant relation branches. The cases include successful OCR-assisted and complementary-evidence predictions as well as failures caused by affective inconsistency, stylized OCR, caption hallucination, and irony.

In the product-related example, production and craftsmanship and consumption experience received the highest prototype activations, and OCR recovered product-related language. In the weak-text example, the image and factual caption provided complementary evidence that was absent from the user-written sentence. In the affectively inconsistent example, the preliminary text and image distributions diverged,

increasing normalized Jensen-Shannon disagreement and reducing direct visual transfer. The OCR failure involved decorative text recognized as a plausible positive expression, whereas the irony-related failure remained difficult because literal text and image evidence were individually plausible while the intended sentiment depended on pragmatic interpretation.

4 Discussion

4.1 Overall Interpretation of the Findings

The results suggest that the main value of SPG-MSA lies not simply in introducing additional information sources, but in controlling *what information is selected, how its reliability is represented, and how relationships among sources are organized before final fusion*. Across both MVSA-Single and TumEmo, SPG-MSA consistently outperformed conventional multimodal baselines and the recent reproduced methods considered in this study. The advantage over earlier fusion architectures was comparatively large, whereas the margin over CLIP-MoE-R was below one macro-F1 percentage point. The latter difference was nevertheless supported by hierarchical bootstrap intervals that did not cross zero.

This distinction is important when interpreting the contribution. Recent CLIP-based and shared-private approaches already learn strong cross-modal representations, and the present results do not indicate a fundamental replacement of such methods. Rather, they show that structured semantic conditioning, local evidence selection, and relation-specific representation can provide an additional and reproducible benefit on top of strong pretrained vision-language features.

This interpretation is also consistent with recent methodological developments. CoDe [16] showed that OCR-derived information and shared-private semantic decomposition can improve sentiment modeling when image and text convey incomplete or discrepant information. SSEDF [18] similarly emphasized the usefulness of separating shared and private semantics, whereas CLIP-MoE [19] demonstrated that adaptive expert specialization can improve multimodal alignment. SPG-MSA extends these directions by distinguishing complementary information from distribution-level disagreement and by introducing a semantic layer that conditions the selection of local evidence. The results therefore support a progressive refinement of multimodal fusion rather than the view that one additional module alone accounts for the observed improvement.

4.2 Why Structured Semantic Conditioning and Fine-Grained Selection Matter

The equal-information experiments provide particularly important evidence for understanding the performance gains. Supplying user text, images, OCR text, and factual captions to a simple concatenation model did not yield performance comparable with SPG-MSA. Even a parameter-matched Transformer using the same four sources

875 remained approximately 1.5 macro-F1 percentage points below the complete architec-
876 ture on both datasets. Thus, the improvement cannot be explained solely by increased
877 information availability or parameter count.

878 This finding is plausible because automatically generated auxiliary sources are
879 heterogeneous in quality. OCR can recover sentiment-relevant signs, product names, or
880 overlaid text, but it can also produce plausible yet incorrect strings. Factual captions
881 reduce the representational distance between visual scenes and language, but they
882 may omit objects or hallucinate content. Unrestricted fusion can therefore amplify
883 redundant or erroneous information. The benefit of auxiliary-quality-aware weighting
884 and local gating is that these sources contribute conditionally rather than uniformly.

885 The ablation results reinforce this interpretation. Fine-grained gating produced
886 one of the largest performance decreases when removed or replaced with a source-
887 level scalar. This observation is consistent with previous fine-grained multimodal
888 studies [20–22], which have shown that sentiment-relevant information is unevenly
889 distributed across words and image regions. SPG-MSA adds a complementary per-
890 spective: local importance is determined not only by cross-modal correspondence but
891 also by semantic-prototype context. In this sense, the prototypes specify *what type of*
892 *evaluative evidence may be relevant*, while the gates determine *where that evidence is*
893 *located within the current sample*.

894

895 4.3 What the Prototype Experiments Reveal

896 The prototype-control experiments clarify both the value and the limits of the pro-
897 posed semantic structure. Random frozen vectors provided almost no benefit over
898 removing prototypes entirely, whereas unnamed learnable vectors performed bet-
899 ter. This pattern indicates that part of the gain arises from introducing adaptive
900 query representations. However, trainable structured prototypes remained stronger
901 than unnamed, sentiment-label, shuffled, and frozen alternatives, suggesting that
902 semantically meaningful initialization also contributes to representation learning.

903 At the same time, the relatively small difference between the complete proto-
904 type inventory and generic brand-attribute prototypes is theoretically informative. It
905 does not support the stronger claim that the selected heritage-inspired dimensions
906 are uniquely suited to general multimodal sentiment analysis. Instead, the evidence
907 supports a more general conclusion: structured evaluative semantic priors are useful,
908 while the optimal inventory is likely to depend on the target domain.

909 The human evaluation leads to a similar interpretation. Prototype Hit@1 values
910 above 0.68 and Hit@3 values above 0.85 indicate substantial correspondence between
911 model rankings and human judgments, but the availability of a “none applicable”
912 category and the lower agreement on TumEmo show that the eight prototypes do
913 not provide an exhaustive semantic taxonomy. TumEmo is organized around basic
914 emotional categories, and many samples—particularly those involving fear, anger,
915 or sadness—can be conveyed primarily through facial or situational cues without a
916 clear connection to temporal continuity, cultural identity, reputation, or consumption
917 experience.

918 Accordingly, the prototype layer should be interpreted as a structured computa-
919 tional prior rather than a validated measurement scale. Future work could replace
920

the fixed inventory with domain-adaptive or hierarchically learned prototypes while retaining semantic anchoring to preserve interpretability.	921
	922
	923
4.4 What Disagreement Modeling Captures—and What It Does Not	924
	925
	926
The controlled mismatch experiments provide a more precise understanding of the disagreement-weighted branch. Removing this branch consistently reduced robustness when text and images produced divergent affective predictions. Its contribution was particularly clear for cross-valence semantic-neighbor mismatches, in which replacement images remained semantically related to the text but conveyed a different affective tendency.	927
	928
	929
	930
	931
	932
This result helps explain why a separate disagreement representation is useful. When semantic similarity remains high, conventional cross-attention may continue to strengthen interaction between the two sources. A distribution-level disagreement signal provides complementary information indicating that semantic correspondence does not necessarily imply affective consistency. This differs from shared-private decomposition alone, which is primarily concerned with common versus modality-specific representations [7, 16, 18].	933
	934
	935
	936
	937
	938
	939
However, same-label semantic mismatches caused a smaller degradation and were less strongly captured by the disagreement mechanism. This is an expected limitation of the normalized Jensen–Shannon score used here. If two semantically unrelated sources nevertheless produce similar sentiment distributions, their disagreement score can remain low. The branch therefore models <i>affective disagreement</i> , not general semantic incompatibility.	940
	941
	942
	943
	944
	945
This distinction is important because disagreement in naturally occurring social-media posts can arise from multiple sources, including sarcasm, irony, irrelevant images, annotation uncertainty, or genuinely contradictory affect. The present experiments cannot identify these causes. A useful extension would therefore combine sentiment-distribution discrepancy with an independent measure of semantic compatibility and evaluate the resulting representation on datasets containing explicit disagreement-type annotations.	946
	947
	948
	949
	950
	951
	952
	953
4.5 Auxiliary Visual Semantics: Useful but Imperfect	954
	955
The OCR and caption analyses further demonstrate why auxiliary visual semantics should be treated as conditional evidence. OCR was retained in only 42.6% of MVSA-Single images and 29.4% of TumEmo images, yet its contribution was larger on MVSA-Single. This pattern is consistent with the greater prevalence of screenshots, signs, and text-bearing images in that dataset. OCR is therefore valuable when available, but the results do not support treating it as a fully populated fourth modality.	956
	957
	958
	959
	960
	961
The caption experiments reveal a different risk. A sentiment-oriented caption prompt produced noticeably stronger caption-only classification, indicating that the frozen caption generator could itself infer affective information. Such a prompt would blur the distinction between visual description and downstream sentiment prediction.	962
	963
	964
	965
	966

967 The factual prompt was therefore the more conservative design. Importantly, remov-
968 ing explicit affective words from factual captions produced only a small additional
969 decrease in the complete model. This suggests that the principal contribution of the
970 caption branch came from object, action, and scene completion rather than direct
971 affective labeling.

972 Nevertheless, factual correctness was not perfect, particularly on TumEmo. Cap-
973 tion hallucinations and important-object omissions remain a source of uncertainty
974 that cosine-based image-caption relevance cannot fully detect. More robust future sys-
975 tems could incorporate entailment-based factuality checks or activate caption-derived
976 evidence only when multiple visual-semantic consistency measures agree.

977

978 4.6 Robustness, Interpretability, and Computational 979 Trade-Offs

980

981 The source-removal experiments show that the model remains primarily text-driven.
982 Removing user text produced the largest performance reduction, especially on
983 TumEmo, whereas removing OCR or captions had a substantially smaller effect.
984 This finding is consistent with the strong Text-CLIP baseline and with the linguis-
985 tic nature of many emotion labels. The contribution of auxiliary visual information
986 should therefore be understood as complementary rather than dominant.

987 Class-level results reveal a similar boundary. Neutral sentiment was the most diffi-
988 cult MVSA-Single category, while fear and bored remained comparatively difficult on
989 TumEmo. These errors are plausible because neutral samples often lack strongly dis-
990 criminative affective cues, while fear, sadness, boredom, and calm can overlap visually
991 or linguistically. The proposed architecture reduces but does not eliminate ambiguity
992 arising from the label structure itself.

993 The explanation-fidelity results provide stronger support for interpretability than
994 attention visualization alone. Fine-grained gates yielded better comprehensiveness and
995 sufficiency than raw cross-attention, and combining gate values with prototype rel-
996 evance produced the strongest perturbation-based attribution. This result supports
997 describing the method as providing *higher-fidelity evidence attribution*, rather than
998 claiming causal explanations. The fact that sufficiency remained above zero further
999 indicates that predictions depend on distributed information across several sources
1000 rather than a single decisive token or patch.

1001 These gains are accompanied by a measurable computational cost. Relative to
1002 MFN, SPG-MSA required more parameters, FLOPs, memory, and online inference
1003 time, while uncached OCR and captioning substantially increased end-to-end latency.
1004 Consequently, the architecture is better suited to offline or near-real-time analytical
1005 settings than to highly constrained streaming applications. This trade-off also qualifies
1006 the comparison with CLIP-MoE-R: when the performance margin is below one macro-
1007 F1 point, deployment decisions should consider whether semantic attribution, OCR
1008 utilization, and disagreement analysis justify the additional complexity.

1009

1010

1011

1012

4.7 Practical Implications, Limitations, and Future Research 1013

The findings suggest several practical uses, but these should be interpreted within the scope of the benchmarks. For multimodal social-media analysis, the framework may be useful when analysts need not only an output label but also information about which visual or textual evidence contributed to the prediction and whether the primary modalities expressed divergent affective tendencies. The quality-regulated treatment of OCR and generated captions is also relevant to systems operating on heterogeneous posts in which auxiliary evidence is intermittently available. 1014
1015
1016
1017
1018
1019
1020

The structured prototype inventory was partly motivated by historically and culturally embedded brand communication. The current results show that these concepts can operate as useful evaluative priors on general datasets, but they do not establish effectiveness for time-honored-brand analysis. Such an application would require dedicated brand identities, multilingual content, domain-specific attributes, and preferably brand-level train-test separation. 1021
1022
1023
1024
1025
1026

More broadly, three limitations define the most productive directions for future work. First, external validity remains constrained by the benchmark construction. MVSA-Single removes directly opposing image-text polarity labels under the adopted preprocessing protocol, while the duplicate-filtered analysis removes affected test observations without rebuilding a fully group-aware training split. Future evaluation should therefore use naturally occurring discrepancy datasets and group-aware resampling procedures. 1027
1028
1029
1030
1031
1032
1033

Second, the semantic representation is structured but not fully identifiable. The eight prototypes overlap conceptually, generic alternatives remain competitive, and disagreement scores capture affective divergence more directly than semantic incompatibility. Adaptive prototype discovery, hierarchical prototype systems, and explicitly annotated disagreement categories could clarify these representational boundaries. 1034
1035
1036
1037
1038

Third, the method depends on computationally expensive auxiliary models. Distillation, shared intermediate representations, conditional activation of OCR and caption branches, and smaller vision-language models could reduce cost without discarding the interpretive benefits demonstrated here. 1039
1040
1041
1042

Taken together, the evidence supports SPG-MSA as a useful extension of structured multimodal fusion, while also showing that its contribution is strongest when semantic interpretation, heterogeneous source quality, and cross-modal disagreement need to be modeled jointly. These observations motivate the more concise conclusions presented below. 1043
1044
1045
1046
1047
1048

5 Conclusion 1049

This study addressed three related challenges in multi-source multimodal sentiment analysis: the absence of an interpretable semantic layer for evidence selection, the uneven reliability of local multimodal information, and the difficulty of distinguishing shared, complementary, and affectively divergent evidence. To address these issues, we developed SPG-MSA, which integrates structured evaluative semantic prototypes, auxiliary-quality-aware fine-grained gating, and relation-specific representation within a unified image-text framework augmented by OCR and factual image captions. 1050
1051
1052
1053
1054
1055
1056
1057
1058

1059 Experiments on MVSA-Single and TumEmo showed that SPG-MSA consistently
1060 outperformed the conventional and recent reproduced baselines considered in this
1061 study. Equal-information and parameter-matched controls further demonstrated that
1062 the observed improvement was not explained solely by additional input sources
1063 or greater model capacity. Prototype controls and human evaluation supported
1064 the usefulness of structured evaluative conditioning, while also showing that the
1065 selected semantic inventory is not universally applicable or uniquely optimal. Con-
1066 trolled perturbations further showed that disagreement-aware representation improves
1067 robustness to affectively inconsistent image-text evidence, although it does not
1068 constitute a general detector of semantic incompatibility.

1069 The study also established an important practical boundary: improved evidence
1070 attribution and robustness come with additional computational and preprocessing
1071 costs. SPG-MSA is therefore most appropriate for offline or near-real-time multimodal
1072 analysis in which semantic interpretation, noisy auxiliary sources, and cross-modal
1073 disagreement are relevant considerations.

1074 Overall, the findings support structured, quality-regulated, relation-aware fusion as
1075 a promising direction for multimodal sentiment analysis while defining clear empirical
1076 boundaries for its interpretation and deployment.

1077

1078 **Author contributions**

1079

1080 L.S. is the sole author and was responsible for the conception and design of the study,
1081 methodology development, software and experimental implementation, data analysis
1082 and validation, visualization, and preparation and revision of the manuscript. The
1083 author reviewed and approved the final manuscript.

1084

1085 **Data availability**

1086

1087 The data analyzed in this study were obtained from the publicly available MVSA-
1088 Single and TumEmo benchmark datasets released by their original authors [28, 29] and
1089 were used in accordance with the applicable data-use terms and licensing conditions.
1090 No newly collected dataset was generated in this study. Where redistribution of the
1091 original social-media content is restricted, the original data are not redistributed by
1092 the author.

1093

1094 **Code availability**

1095

1096 The code developed and used for the analyses in this study is provided as Supplemen-
1097 tary Material accompanying this article.

1098

1099 **Ethics statement**

1100

1101 The present study used publicly available benchmark datasets containing previously
1102 released multimodal social-media data. No new participants were recruited, no addi-
1103 tional human-participant data were collected, and the authors did not have access
1104

to re-identification keys or use direct personal identifiers as model inputs. User-identifying information was excluded from the analyses and qualitative examples. The secondary analysis was reviewed by the Ethics Committee of Shandong Agriculture and Engineering University and was determined to be exempt from further ethical review.

Funding

The author received no specific funding for this work.

Competing interests

The author declares no competing interests.

References

- [1] Park, J., Hong, E. & Park, Y. Toward a new business model of retail industry: The role of brand experience and brand authenticity. *Journal of Retailing and Consumer Services* **74**, 103426 (2023).
- [2] Zhang, Y. & Guo, X. “New and old”: Consumer evaluations of co-branding between new brands and chinese time-honored brands. *Journal of Retailing and Consumer Services* **73**, 103293 (2023).
- [3] Yin, C.-C., Tang, Y.-C., Chiu, H.-C., Hsieh, Y.-C. & Lai, Y.-T. Telling an authentic story by aligning with your product type and price. *Journal of Business Research* **161**, 113799 (2023).
- [4] Gan, C. *et al.* A multimodal fusion network with attention mechanisms for visual-textual sentiment analysis. *Expert Systems with Applications* **242**, 122731 (2024).
- [5] Wei, Y. *et al.* Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection (2023). Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, pp. 5240–5252.
- [6] Zhu, C. *et al.* SKEAFN: Sentiment knowledge enhanced attention fusion network for multimodal sentiment analysis. *Information Fusion* **100**, 101958 (2023).
- [7] Zhao, X., Chen, Y., Liu, S. & Tang, B. Shared-private memory networks for multimodal sentiment analysis. *IEEE Transactions on Affective Computing* **14**, 2889–2900 (2023).
- [8] Liu, H. *et al.* Social image-text sentiment classification with cross-modal consistency and knowledge distillation. *IEEE Transactions on Affective Computing* **14**, 3332–3344 (2023).

1151 [9] Wang, H., Ren, C. & Yu, Z. Multimodal sentiment analysis based on cross-
1152 instance graph neural networks. *Applied Intelligence* **54**, 3403–3416 (2024).
1153

1154 [10] Gong, L., He, X. & Yang, J. An image-text sentiment analysis method using multi-
1155 channel multi-modal joint learning. *Applied Artificial Intelligence* **38**, 2371712
1156 (2024).

1157 [11] Lin, R. & Hu, H. MissModal: Increasing robustness to missing modality in mul-
1158 timodal sentiment analysis. *Transactions of the Association for Computational*
1159 *Linguistics* **11**, 1686–1702 (2023).
1160

1161 [12] Zhao, X., Poria, S., Li, X., Chen, Y. & Tang, B. Toward robust multimodal
1162 sentiment analysis using multimodal foundational models. *Expert Systems with*
1163 *Applications* **276**, 126974 (2025).
1164

1165 [13] Zhao, X., Li, X., Jiang, R. & Tang, B. Resolving multimodal ambiguity via
1166 knowledge-injection and ambiguity learning for multimodal sentiment analysis.
1167 *Information Fusion* **115**, 102745 (2025).
1168

1169 [14] Li, Z. *et al.* Hierarchical denoising representation disentanglement and dual-
1170 channel cross-modal-context interaction for multimodal sentiment analysis.
1171 *Expert Systems with Applications* **252**, 124236 (2024).
1172

1173 [15] Yang, B. & Li, J. Visual elements mining as prompts for instruction learning
1174 for target-oriented multimodal sentiment classification (2023). Findings of the
1175 Association for Computational Linguistics: EMNLP 2023, pp. 6062–6075.
1176

1177 [16] Wu, D., Yang, D., Shen, H., Ma, C. & Zhou, Y. Resolving sentiment discrepancy
1178 for multimodal sentiment detection via semantics completion and decomposition.
1179 *Pattern Recognition* **172**, 112719 (2026).

1180 [17] Zhang, B., Wang, J., Sun, B. & Ma, N. Multi-level global and local fusion for
1181 multimodal sentiment analysis. *Memetic Computing* **18**, 36 (2026).
1182

1183 [18] Zhang, W., Du, Y., Wang, H., Liu, J. & Li, X. SSEDf: A shared-private semantic
1184 enhanced dynamic fusion network for multimodal sentiment analysis. *Expert*
1185 *Systems with Applications* **303**, 130667 (2026).
1186

1187 [19] Wang, J. *et al.* Fine-tuning CLIP with mixture of experts and cross-modal align-
1188 ment via contrastive learning for multimodal sentiment analysis. *Neurocomputing*
1189 **670**, 132417 (2026).
1190

1191 [20] Huang, C. *et al.* TeFNA: Text-centered fusion network with crossmodal atten-
1192 tion for multimodal sentiment analysis. *Knowledge-Based Systems* **269**, 110502
1193 (2023).
1194
1195
1196

[21] Chen, D., Su, W., Wu, P. & Hua, B. Joint multimodal sentiment analysis based on information relevance. <i>Information Processing & Management</i> 60 , 103193 (2023).	1197 1198 1199 1200
[22] Xiao, L. <i>et al.</i> Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. <i>Information Processing & Management</i> 60 , 103508 (2023).	1201 1202 1203 1204
[23] Cui, C. <i>et al.</i> Paddleocr 3.0 technical report. <i>arXiv preprint arXiv:2507.05595</i> (2025).	1205 1206 1207
[24] Li, J., Li, D., Savarese, S. & Hoi, S. C. H. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023). Proceedings of the 40th International Conference on Machine Learning, pp. 19730–19742, PMLR.	1208 1209 1210 1211
[25] Radford, A. <i>et al.</i> Learning transferable visual models from natural language supervision (2021). Proceedings of the 38th International Conference on Machine Learning, pp. 8748–8763, PMLR.	1212 1213 1214 1215
[26] Lin, J. Divergence measures based on the shannon entropy. <i>IEEE Transactions on Information Theory</i> 37 , 145–151 (1991).	1216 1217 1218
[27] Khosla, P. <i>et al.</i> Supervised contrastive learning (2020). Advances in Neural Information Processing Systems, pp. 18661–18673, Curran Associates, Inc.	1219 1220 1221
[28] Niu, T., Zhu, S., Pang, L. & El Saddik, A. Sentiment analysis on multi-view social data (2016). MultiMedia Modeling, pp. 15–27, Springer.	1222 1223 1224
[29] Yang, X., Feng, S., Zhang, Y. & Wang, D. Multimodal sentiment detection based on multi-channel graph neural networks (2021). Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 328–339.	1225 1226 1227 1228
[30] Reimers, N. & Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks (2019). Proceedings of EMNLP-IJCNLP 2019, pp. 3982–3992, Association for Computational Linguistics.	1229 1230 1231 1232
[31] Zauner, C. <i>Implementation and benchmarking of perceptual image hash functions</i> . Master’s thesis, Upper Austria University of Applied Sciences (2010).	1233 1234 1235
[32] Loshchilov, I. & Hutter, F. Decoupled weight decay regularization (2019). International Conference on Learning Representations, OpenReview.net.	1236 1237 1238
[33] Mohammad, S. M. & Turney, P. D. Crowdsourcing a word-emotion association lexicon. <i>Computational Intelligence</i> 29 , 436–465 (2013).	1239 1240 1241 1242

1243 [34] Krippendorff, K. *Content Analysis: An Introduction to Its Methodology* 4 edn
1244 (Sage, 2018).
1245
1246 [35] Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural
1247 networks (2017). Proceedings of the 34th International Conference on Machine
1248 Learning, pp. 1321–1330, PMLR.
1249
1250 [36] DeYoung, J. *et al.* ERASER: A benchmark to evaluate rationalized nlp mod-
1251 els (2020). Proceedings of the 58th Annual Meeting of the Association for
1252 Computational Linguistics, pp. 4443–4458.
1253
1254 [37] McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation
1255 and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
1256
1257 [38] Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman &
1258 Hall/CRC, 1993).
1259
1260 [39] Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian*
1261 *Journal of Statistics* **6**, 65–70 (1979).
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288

Supplementary Materials

This section contains the supplementary tables and reproducibility details cited throughout the main manuscript. The numbering is independent of the main article.
Supplementary table note. Supplementary Table [S1](#) provides the detailed feature-level comparison that supports the methodological distinctions summarized in the Introduction.

1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334

1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380

Panel A. Input and conditioning mechanisms

Table S1 Conceptual comparison with closely related multimodal sentiment methods.

Method	OCR	Generated auxiliary source	Semantic prototypes	Token/patch gate
CoDe	Yes	No	No	No
DualScope	No	Generated text/image	No	Hierarchical
SSEDF	No	No	No	Source-level
CLIP-MoE	No	No	No	Expert-level
SPG-MSA	Yes	Factual caption	Yes	Yes

Panel B. Representation and reliability mechanisms

Method	Shared information	Complementary information	Distribution disagreement	Auxiliary-quality signal
CoDe	Yes	Private representation	Contrastive discrepancy	No
DualScope	Yes	Global-local information	Implicit	No
SSEDF	Yes	Private semantics	Implicit	No
CLIP-MoE	Implicit	Expert-specific information	Logit modulation	No
SPG-MSA	Yes	Yes	Normalized JS divergence	Yes

Supplementary table note. Supplementary Table S2 lists the operational definitions and representative evidence used to construct the eight structured evaluative semantic prototypes.

Table S2 Structured evaluative semantic-prototype inventory.

Prototype	Operational definition	Representative evidence
Temporal continuity	Long-term identity, history, duration, or persistence	Historical narratives, establishment dates, old locations
Cultural identity	Regionally, socially, or culturally embedded meanings	Local symbols, traditional motifs, collective identity
Production and craftsmanship	Skilled, process-oriented, handmade, or quality-related production	Tools, production steps, workmanship, materials
Perceived authenticity	Consistency among claims, identity, origin, and observable evidence	Provenance, genuineness, credible narratives
Nostalgia and identification	Memory-related affect or identification with past experience	Retro design, childhood memories, old photographs
Trust and reputation	Confidence based on stable performance, recognition, or accumulated standing	Recommendations, awards, commitments, reputation
Contemporary renewal	Adaptation, redesign, collaboration, or modern reinterpretation	New packaging, digital services, co-branding
Consumption experience	Evaluation of products, services, price, environment, or use	Taste, queues, convenience, service, value

Supplementary table note. Supplementary Table S3 reports the full class distribution of the fixed training, validation, and test partitions used in all controlled experiments.

Table S3 Class distribution in the fixed study splits.

Dataset	Class	Training	Validation	Test	Total
MVSA-Single	Positive	2,146	268	269	2,683
	Neutral	376	47	47	470
	Negative	1,086	136	136	1,358
	All	3,608	451	452	4,511
TumEmo	Angry	11,635	1,454	1,456	14,545
	Bored	25,826	3,228	3,231	32,285
	Calm	14,487	1,811	1,811	18,109
	Fear	16,211	2,026	2,028	20,265
	Happy	40,214	5,027	5,029	50,270
	Love	27,609	3,451	3,453	34,513
	Sad	20,222	2,528	2,528	25,278
	All	156,204	19,525	19,536	195,265

Supplementary table note. Supplementary Table S4 gives the complete five-seed performance comparison for conventional and recent reproduced baselines.

Table S4 Overall controlled performance on MVSA-Single and TumEmo. Results are mean $\pm$ standard deviation over five random seeds.

Model	MVSA ACC	MVSA Macro-F1	TumEmo ACC	TumEmo Macro-F1
Image-CLIP	0.6558 ± 0.0064	0.6405 ± 0.0081	0.4827 ± 0.0035	0.3561 ± 0.0058
Text-CLIP	0.7046 ± 0.0059	0.6961 ± 0.0067	0.6196 ± 0.0029	0.6128 ± 0.0034
Early Fusion	0.7137 ± 0.0061	0.7049 ± 0.0069	0.5989 ± 0.0037	0.5848 ± 0.0041
MVCN	0.7568 ± 0.0048	0.7497 ± 0.0055	0.6507 ± 0.0028	0.6442 ± 0.0032
SKEAFN-IT	0.7641 ± 0.0044	0.7562 ± 0.0051	0.6638 ± 0.0026	0.6579 ± 0.0029
SPMN-IT	0.7685 ± 0.0042	0.7609 ± 0.0049	0.6689 ± 0.0025	0.6631 ± 0.0028
CIGNN	0.7511 ± 0.0050	0.7333 ± 0.0057	0.6738 ± 0.0023	0.6706 ± 0.0025
MMJL	0.7698 ± 0.0040	0.7623 ± 0.0046	0.6713 ± 0.0027	0.6670 ± 0.0030
MFN	0.7746 ± 0.0039	0.7668 ± 0.0045	0.6795 ± 0.0022	0.6744 ± 0.0026
DualScope-R	0.8049 ± 0.0041	0.7967 ± 0.0048	–	–
CoDe-R	0.8096 ± 0.0037	0.8018 ± 0.0044	0.7148 ± 0.0022	0.7086 ± 0.0026
SSEDF-R	0.8118 ± 0.0035	0.8049 ± 0.0041	0.7169 ± 0.0020	0.7114 ± 0.0024
CLIP-MoE-R	0.8163 ± 0.0033	0.8088 ± 0.0039	0.7207 ± 0.0018	0.7156 ± 0.0022
SPG-MSA	0.8247 ± 0.0035	0.8172 ± 0.0041	0.7296 ± 0.0019	0.7241 ± 0.0023

1473
1474
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518

Supplementary table note. Supplementary Table S5 reports both equal-information controls and component-wise ablations used to separate gains from additional inputs, parameter count, and the proposed modules.

Table S5 Equal-information controls and component ablation.

Configuration	MVSA Macro-F1	TumEmo Macro-F1
<i>Equal-information controls</i>		
Four-source concatenation	0.7769	0.6830
Four-source Transformer	0.7902	0.6971
Four-source cross-attention	0.7978	0.7035
Parameter-matched Transformer	0.8024	0.7086
SPG-MSA	0.8172	0.7241
<i>Component ablation</i>		
Without semantic prototypes	0.7965	0.7044
Without fine-grained gating	0.7938	0.7019
Gate replaced by source scalar	0.7907	0.6994
Without shared-information branch	0.8056	0.7125
Without complementary-information branch	0.8031	0.7103
Without disagreement-weighted branch	0.8017	0.7082
Without centered cross-covariance loss	0.8084	0.7128
Without alignment loss	0.8053	0.7101
Without prototype anchoring	0.8098	0.7144
Without prototype diversity	0.8091	0.7139
Without gate-activation constraint	0.8112	0.7185

Supplementary table note. Supplementary Table S6 reports computational prototype controls together with the human relevance evaluation.

Table S6 Semantic-prototype computational controls and human relevance evaluation.

Prototype configuration	MVSA Macro-F1	TumEmo Macro-F1
No prototypes	0.7965	0.7044
Random frozen vectors	0.7981	0.7052
Unnamed learnable vectors	0.8036	0.7101
Sentiment-label prototypes	0.8078	0.7150
Generic brand-attribute prototypes	0.8102	0.7173
Shuffled semantic names	0.8011	0.7078
Frozen structured prototypes	0.8129	0.7201
Trainable structured prototypes	0.8172	0.7241

Dataset	Krippendorff's α	None applicable	Hit@1	Hit@3	mAP
MVSA-Single	0.781	6.5%	0.714	0.892	0.704
TumEmo	0.742	9.0%	0.681	0.851	0.663

Supplementary table note. Supplementary Table [S7](#) summarizes auxiliary-source contributions, OCR availability, caption audit results, and the caption-prompt leakage control.

1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656

Table S7 Auxiliary-source contribution, OCR characteristics, and caption audit.

Source configuration	MVSA Macro-F1	TumEmo Macro-F1
User text + image	0.7921	0.6980
+ OCR	0.8076	0.7147
+ factual caption	0.8104	0.7161
+ OCR + factual caption	0.8172	0.7241
Full sources without auxiliary-quality signals	0.8089	0.7138

OCR measure	MVSA-Single	TumEmo
Images containing retained OCR text	42.6%	29.4%
Mean boxes per OCR-positive image	2.8	2.1
Mean recognition confidence	0.824	0.806
Images represented by [NO_OCR]	57.4%	70.6%
OCR samples containing sentiment-related words	11.8%	8.5%

Caption audit measure	MVSA-Single	TumEmo
Captions containing explicit affective words	2.8%	3.4%
Mean caption length (tokens)	16.7	15.9
Mean image-caption similarity	0.742	0.716
Empty or invalid captions	0.6%	0.9%
Human factual-correctness rate	88.7%	84.9%
Hallucinated-object rate	5.7%	8.3%
Important-object omission rate	12.3%	16.0%

Caption setting	MVSA F1 (caption-only / full)	TumEmo F1 (caption-only / full)
Sentiment-oriented prompt	0.6235 / 0.8209	0.5161 / 0.7275
Factual prompt	0.5918 / 0.8172	0.4860 / 0.7241
Factual caption with affective words removed	0.5867 / 0.8159	0.4803 / 0.7228
No caption	– / 0.8076	– / 0.7147

Supplementary table note. Supplementary Table [S8](#) provides the complete missing-source and controlled-mismatch robustness results discussed in the main Results section.

1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702

1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748

Table S8 Robustness under unavailable and perturbed sources. Changes are absolute accuracy differences from complete input.

Condition	MVSA ACC	Change	TumEmo ACC	Change
Complete input	0.8247	–	0.7296	–
OCR unavailable	0.8152	-0.0095	0.7214	-0.0082
Caption unavailable	0.8176	-0.0071	0.7230	-0.0066
Image unavailable	0.7559	-0.0688	0.6601	-0.0695
User text unavailable	0.7168	-0.1079	0.5726	-0.1570
20% random cross-pair mismatch	0.7833	-0.0414	0.6915	-0.0381
20% OCR-token corruption	0.8141	-0.0106	0.7198	-0.0098
20% sentiment-word masking	0.7862	-0.0385	0.6790	-0.0506
Random mismatch without disagreement branch	0.7584	-0.0663	0.6672	-0.0624
Same-label semantic mismatch	0.7958	-0.0289	0.7037	-0.0259
Cross-valence semantic-neighbor mismatch	0.7619	-0.0628	0.6754	-0.0542

Supplementary table note. Supplementary Table S9 reports perturbation-based explanation fidelity together with model-size, memory, latency, and preprocessing costs.	1749
	1750
	1751
	1752
	1753
	1754
	1755
	1756
	1757
	1758
	1759
	1760
	1761
	1762
	1763
	1764
	1765
	1766
	1767
	1768
	1769
	1770
	1771
	1772
	1773
	1774
	1775
	1776
	1777
	1778
	1779
	1780
	1781
	1782
	1783
	1784
	1785
	1786
	1787
	1788
	1789
	1790
	1791
	1792
	1793
	1794

1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840

Table S9 Perturbation-based explanation fidelity and computational efficiency.

Explanation method	MVSA Comp. $\uparrow$	MVSA Suff. $\downarrow$	TumEmo Comp. $\uparrow$	TumEmo Suff. $\downarrow$
Random evidence	0.041	0.184	0.036	0.197
Raw cross-attention	0.092	0.127	0.084	0.139
Gradient attribution	0.118	0.104	0.109	0.118
Fine-grained gate	0.143	0.083	0.132	0.096
Gate + prototype attribution	0.158	0.071	0.146	0.083

Model	Trainable parameters	FLOPs/sample	Peak VRAM	Time/epoch	Online latency	Offline preprocessing
Text-CLIP	63.1 M	5.8 G	4.6 GB	42 s	4.8 ms	None
Early Fusion	149.6 M	17.7 G	7.1 GB	58 s	9.6 ms	None
MFN	158.3 M	20.4 G	8.3 GB	75 s	12.8 ms	None
SPG-MSA	171.9 M	25.9 G	10.3 GB	98 s	18.7 ms	1.15 s
OCR component	–	–	–	–	–	0.31 s
Caption component	–	–	–	–	–	0.84 s

Supplementary table note. Supplementary Table [S10](#) provides the within-dataset duplicate counts used in the duplicate audit.

1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886

1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932

Table S10 Within-dataset duplicate observations. Categories are not mutually exclusive; an observation may be counted in more than one category.

Dataset	Exact-text observations	Near-text observations	Exact-image observations	Near-image observations
MVSA-Single	41	88	23	51
TumEmo	3,184	7,462	1,376	4,805

Supplementary table note. Supplementary Table [S11](#) reports sensitivity to removing test observations associated with cross-split duplicate groups; models were not retrained for this analysis.

1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978

1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024

Table S11 Sensitivity to cross-split duplicate removal from the test set. Models were not retrained for this sensitivity analysis.

Dataset	Groups	Affected test	ACC orig.	ACC filt.	Macro-F1 orig.	Macro-F1 filt.
MVSA-Single	17	31	0.8247	0.8211	0.8172	0.8134
TumEmo	1,294	2,217	0.7296	0.7281	0.7241	0.7227

Supplementary table note. Supplementary Table S12 reports class-level precision, recall, F1, variability, and test support for both benchmarks.

Table S12 Class-level precision, recall, F1, and test-set support. Metrics are mean $\pm$ standard deviation across five independently trained models.

Dataset	Class	Support	Precision	Recall	F1
MVSA-Single	Positive	269	0.884 ± 0.004	0.861 ± 0.006	0.872 ± 0.005
	Neutral	47	0.747 ± 0.010	0.784 ± 0.011	0.765 ± 0.009
	Negative	136	0.810 ± 0.007	0.821 ± 0.008	0.815 ± 0.007
TumEmo	Angry	1,456	0.741 ± 0.004	0.729 ± 0.005	0.735 ± 0.004
	Bored	3,231	0.661 ± 0.006	0.635 ± 0.007	0.648 ± 0.006
	Calm	1,811	0.751 ± 0.004	0.733 ± 0.005	0.742 ± 0.004
	Fear	2,028	0.644 ± 0.006	0.619 ± 0.007	0.631 ± 0.006
	Happy	5,029	0.803 ± 0.003	0.783 ± 0.003	0.793 ± 0.003
	Love	3,453	0.789 ± 0.003	0.773 ± 0.004	0.781 ± 0.003
	Sad	2,528	0.731 ± 0.004	0.747 ± 0.004	0.739 ± 0.004

Supplementary table note. Supplementary Table S13 provides the numerical values underlying the corruption-ratio curves in Figure 6.

2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116

Table S13 Accuracy across increasing corruption ratios.

Dataset	Perturbation	0%	10%	20%	30%	40%	50%
MVSA	Random image mismatch	0.8247	0.8045	0.7833	0.7632	0.7419	0.7210
	OCR corruption	0.8247	0.8195	0.8141	0.8072	0.7994	0.7906
	Sentiment masking	0.8247	0.8060	0.7862	0.7650	0.7418	0.7165
	Mismatch without disagreement	0.8092	0.7830	0.7584	0.7335	0.7042	0.6760
TumEmo	Random image mismatch	0.7296	0.7112	0.6915	0.6718	0.6500	0.6285
	OCR corruption	0.7296	0.7245	0.7198	0.7130	0.7050	0.6962
	Sentiment masking	0.7296	0.7041	0.6790	0.6532	0.6265	0.5980
	Mismatch without disagreement	0.7143	0.6910	0.6672	0.6418	0.6130	0.5821

Supplementary table note. Supplementary Table S14 reports sensitivity to the number of semantic prototypes and the gate temperature used in the implementation.

Table S14 Sensitivity to semantic-prototype number and gate temperature.

Setting	MVSA Macro-F1	TumEmo Macro-F1
4 prototypes	0.8037	0.7116
6 prototypes	0.8129	0.7190
8 prototypes	0.8172	0.7241
10 prototypes	0.8158	0.7229
12 prototypes	0.8121	0.7197
$\tau = 0.03$	0.8094	0.7170
$\tau = 0.05$	0.8152	0.7220
$\tau = 0.07$	0.8172	0.7241
$\tau = 0.10$	0.8148	0.7218
$\tau = 0.20$	0.8087	0.7154