Exploring Cognitive and Aesthetic Causality for Multimodal Aspect-Based Sentiment Analysis

Luwei Xiao, *Student Member, IEEE*, Rui Mao*, *Member, IEEE*, Shuai Zhao, Qika Lin, Yanhao Jia, Liang He, and Erik Cambria, *Fellow, IEEE*

Abstract—Multimodal aspect-based sentiment classification (MASC) is an emerging task due to an increase in user-generated multimodal content on social platforms, aimed at predicting sentiment polarity toward specific aspect targets (i.e., entities or attributes explicitly mentioned in text-image pairs). Despite extensive efforts and significant achievements in existing MASC, substantial gaps remain in understanding fine-grained visual content and the cognitive rationales derived from semantic content and impressions (cognitive interpretations of emotions evoked by image content). In this study, we present Chimera: a **c**ognitive and aest**h**et**i**c senti**me**nt causality unde**r**st**a**nding framework to derive fine-grained holistic features of aspects and infer the fundamental drivers of sentiment expression from both semantic perspectives and affective-cognitive resonance (the synergistic effect between emotional responses and cognitive interpretations). The framework aligns visual patches with words, extracts coarse and fine-grained visual features, translates them into textual descriptions, and uses LLM-generated sentimental causes and impressions to boost sensitivity to affective cues. Experiments on MASC datasets show the model's effectiveness and greater flexibility compared to LLMs like GPT-4o. We have publicly released the complete implementation and dataset at https://github.com/SenticNet/Chimera

Index Terms—Multimodal aspect-based sentiment classification, Sentiment causality, Large language models, Affective-cognitive resonance.

1 Introduction

Multimodal aspect-based sentiment classification (MASC) is a valuable task for analyzing user-generated multimodal content on social platforms, aiming to predict the sentiment polarity of a specific target/aspect term within a sentence, based on an image-text pair. In an era marked by growing global interconnectedness, social platforms have become essential channels for individuals to express opinions and share experiences [1]–[4]. These platforms support multimodal content, blending text and visual media, which better reflects how sentiment is conveyed [5]. Consequently, analyzing fine-grained sentiment expression in multimodal scenarios not only improves the depth of sentiment classification but also aligns with the natural manner in which users express opinions and emotions, ultimately supporting more accurate sentiment analysis for applications in finance [6], [7], social research [8]–[10], and human-computer interaction [11], [12]. Current methodologies for MASC can be broadly divided into two principal categories: visual-text fusion-based approaches and translation-based approaches. Visual-text fusion-based methods address MASC by directly integrating visual content with textual features through various attention-based mechanisms [13]–[18].

Yu *et al.* [13] were the first to propose the utilization of ResNet for image feature extraction in conjunction with BERT for language sequence modeling, subsequently feeding these components into a BERT encoder to facilitate the interactive modeling of cross-modal representations. Ling *et al.* [16] introduced a vision-language pre-training framework that leverages Faster R-CNN for extracting object-level visual features and BART for generating textual features, with the model pre-trained using three task-specific strategies targeting the language, vision, respectively. Yu *et al.* [15] presented a novel multi-task learning framework Image-Target Matching Network (ITM), which concurrently performs coarse-to-fine-grained visual-textual relevance detection and visual object-target alignment through cross-modal Transformers.

Translation-based approaches focus on mapping visual content into the language space as auxiliary textual representations, leveraging this supplementary information, or integrating it with visual features to enhance MASC [19]–[24]. Khan *et al.* [19] translated the image into a corresponding caption, which is then jointly input with the sentence into BERT to predict the sentiment polarity associated with specific targets. Yang *et al.* [25] exploit a face-sensitive, translation-based approach that translates facial expressions in images into textual sentiment cues, which are then selectively aligned and fused with the targets for enhanced sentiment analysis. Xiao *et al.* [21] proposed the CoolNet framework, which generates visual captions for images and extracts syntactic and semantic features from the textual modality, subsequently fusing these with visual features through a cross-modal Transformer.

- *Luwei Xiao, and Liang He are with the School of Computer Science and Technology, East China Normal University, Shanghai 200062, China. E-mail: louisshaw@stu.ecnu.edu.cn, lhe@cs.ecnu.edu.cn*
- *Rui Mao, Shuai Zhao, Yanhao Jia and Erik Cambria are with the College of Computing and Data Science, Nanyang Technological University, Singapore 639798. E-mail:{rui.mao, shuai.zhao, cambria}@ntu.edu.sg, yanhao002@e.ntu.edu.sg*
- *Qika Lin is with the Saw Swee Hock School of Public Health, National University of Singapore 119077. E-mail: linqika@nus.edu.sg*

** Corresponding author: Rui Mao*

Despite substantial efforts and promising advancements, current solutions continue to encounter the following challenges. First, excessive duplicative visual patches can overshadow critical visual clues relevant to the specific target, leading to considerable misalignment during patch-token interactions. These small visual patches often lack semantic coherence compared to complete visual regions, particularly when aligning targets with their corresponding objects in an image, potentially leading to ambiguous semantic representations. Second, limited studies have focused on the underlying rationale behind sentiment cues, particularly from the perspectives of semantic content and affective-cognitive resonance. Owing to the multimodal nature of Twitter content, which spans diverse facets of daily life, inferring the sentiment associated with specific targets necessitates not only an understanding of the surface-level information in text and images (e.g., facial expressions) but also an in-depth comprehension of the contextual background of particular events and the impressions evoked by the image's content and aesthetic attributes.

To address the aforementioned challenges, this paper proposes Chimera: a **c**ognitive and aest**h**et**i**c senti**me**nt causality unde**r**st**a**nding framework. This framework aims to incorporate and align fine-grained features of specific targets and reasons about semantic and impression rationales. However, two critical issues must be resolved to achieve these objectives: 1) How can specific targets in a sentence be aligned with their corresponding object-level fine-grained features in an image? 2) How can the model be enabled to reason about the emotional causal reasons within the semantic content of image-text pairs and the affective resonance evoked by image aesthetic attributes? For the first question, we propose to make the cross-modal alignment of the target via the visual patch-level by linguistic-aware patch-token alignment and object-level by accurately translating the object feature into language space. Regarding the second issue, while a recent study [26] developed a reasoning dataset for MASC, this dataset primarily explains the emotional causes within textual content and lacks reasoning capabilities for visual content and the affective resonance evoked by images, limiting its suitability for the multimodal nature of this task. Consequently, we employ a large language model (LLM), GPT-4o, to generate the semantic rationale and impression rationale to understand the causal foundations of emotions.

Specifically, our proposed framework first extracts visual patch-level and textual features, feeding them into a tailored linguistic-aware patch-token alignment (LPA) module to achieve patch-token alignment. Concurrently, a translation module (TM) translates the holistic image or object-level content into aesthetic captions or facial descriptions, leveraging multimodal named entity annotations from the work of Wang *et al.* [27]. The TM-generated text, along with the sentence and aspect, is then input into a generative module for multi-task learning to produce sentiment polarity, semantic rationale (SR), and impression rationale (IR). By bootstrapping the model's perception of underlying rationale through an in-depth understanding of textual and visual content as well as the affective resonance evoked by images, it enhances the performance of sentiment classification.

In a nutshell, the primary contributions are as follows:

- We propose a novel framework for MASC that aligns specific targets with their corresponding visual objects at the patch-token and object levels while equipping the model with causal rationale reasoning ability for semantic rationale (SR), and impression rationale (IR).
- We approach this task by enabling the model to grasp the semantic content of image-text pairs and the affective resonance evoked by images. To our knowledge, we are the first to collect semantic and impression rationale data for the MASC task, based on existing MASC datasets, extending its content to incorporate semantic and impression rationale, offering a valuable resource for advancing MASC research.
- Experiments on three widely-used Twitter benchmarks demonstrate that our proposed method outperforms previous approaches, achieving state-of-the-art performance. Further evaluations validate the effectiveness of our approach for MASC tasks.

The remainder of this paper is organized as follows: Section 2 provides an overview of related research on multimodal aspect-based sentiment classification, image aesthetic assessment, and multimodal learning. Section 3 details the proposed framework, including linguistics-aware patch-token alignment, the translation-based module, causal rationale dataset construction, and LLM-based annotation generation. Main experimental results are presented in Section 4, and the in-depth analysis is shown in 5, followed by conclusions in Section 6.

2 Related Work

This section reviews key methods in multimodal aspect-based sentiment analysis and image aesthetic assessment. Additionally, as our novel rationale dataset is constructed using an LLM, we introduce LLMs for data annotation.

2.1 Multimodal Aspect-based Sentiment Analysis

Sentiment analysis is a well-established research area focused on understanding and identifying human emotions and opinions across various contexts [28]–[33]. With the exponential growth of user-generated multimodal content (e.g., image-text pairs, video clips) on social platforms [34]–[37] has drawn substantial attention to Multimodal Aspect-based Sentiment Analysis (MABSA) [38]–[42]. The MABSA task consists of two sub-tasks: Multimodal Aspect Term Extraction (MATE) and our focused MASC task. MATE [43] is essentially a named entity recognition task aimed at identifying all relevant specific targets within the textual content of an image-text pair. MASC [44], [45] is a text classification task in which specific targets are provided, requiring the identification of their sentiment polarity (positive, neutral, or negative) based on the given image-text pair. A series of recent studies have successfully unified these two subtasks into a single framework, effectively streamlining the MABSA process [16], [17], [24], [46]–[49].

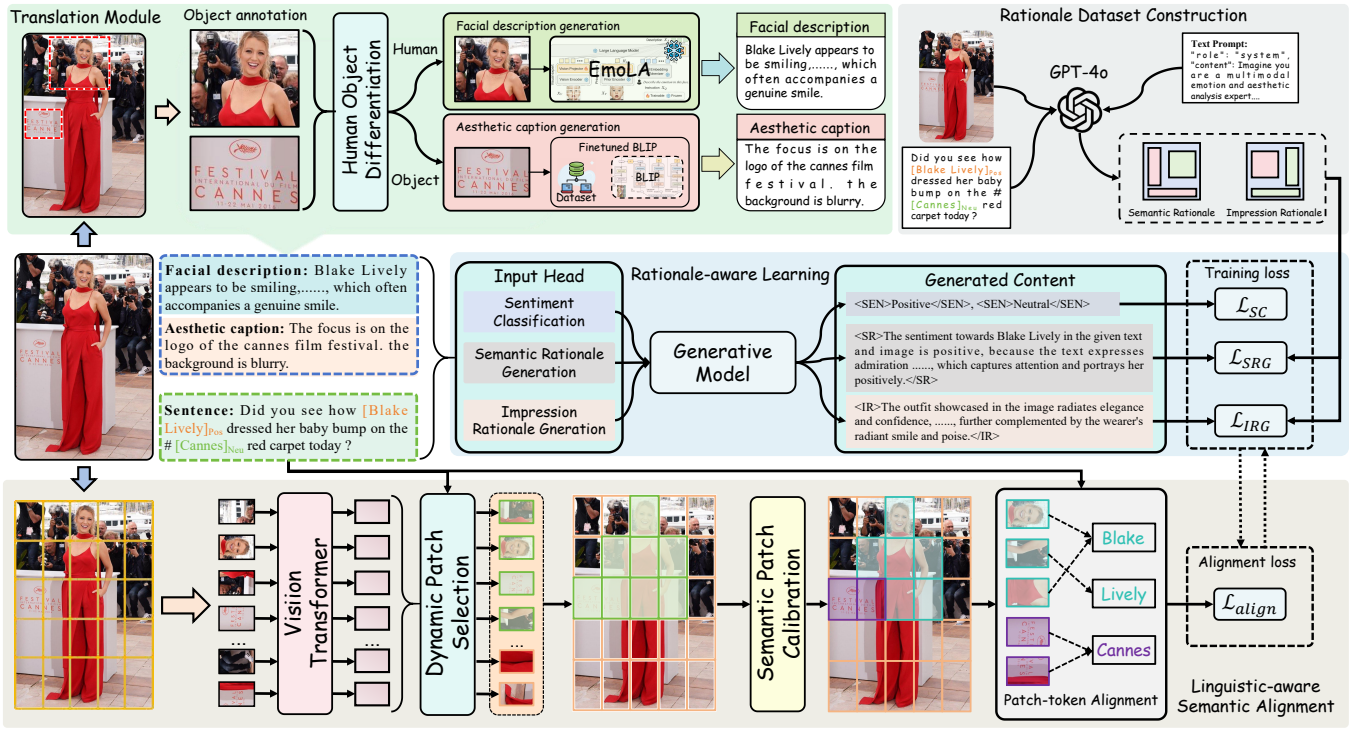

Fig. 1. The overall framework of the proposed Chimera. Chimera consists of four parts: Translation Module, Rationale Dataset Construction, Linguistic-aware Semantic Alignment, and Rationale-Aware Learning.

Among these studies, Yu *et al.* [14] proposed the Entity-Sensitive Attention and Fusion Network (ESAFN), which employs entity-oriented attention combined with a visual gate mechanism to model entity-sensitive inter-dynamics for MASC. Ju *et al.* [46] were the first to integrate MATE and MASC into a end-to-end task, developing a joint learning framework with cross-modal relation detection. Kruk *et al.* [37] proposed a multimodal framework for Instagram intent detection, integrating three taxonomies and the MDID dataset. It demonstrates that text-image fusion enhances accuracy by 9.6% under semiotic divergence, emphasizing the necessity of multimodal models for capturing the non-intersective "meaning multiplication" inherent in social media. Yang *et al.* [17] improved cross-modal alignment modeling through a Transformer-based multi-task learning framework, incorporating text-guided cross-modal interactions and using adjective–noun pairs as supervision for a visual auxiliary task.

Zhou *et al.* [18] developed an aspect-oriented multimodal fusion approach that constructs an informative dependency graph to minimize additional visual and textual noise in cross-modal interactions by selectively processing aspect-relevant textual and image features. Huang *et al.* [22] put forward to mapping images into scene graphs, using triplet semantic relationships among entities along with image captions to construct a relatedness matrix for achieving cross-modal alignment in MASC. More recently, Xiao *et al.* [24] introduced the Atlantis, a trident-shaped architecture that incorporates aesthetic attributes to enhance the emotional resonance of visual content. Fan *et al.* [26] devised a Flan-T5-based multi-task learning architecture to enhance the model's reasoning capabilities for inferring underlying and direct causes of sentiment expressions. Additionally, they constructed a practical causal dataset for MASC. Our proposed method aims to achieve cross-modal alignment at the patch and object levels while equipping the model with reasoning capabilities to discern the semantic and impression rationale underlying sentiment expressions.

2.2 Image Aesthetic Assessment

Image aesthetics play a fundamental role in shaping viewers' emotional responses and overall aesthetic experience through complex psychological and cognitive processes [50]. Image aesthetics pertain to the subjective evaluation and appreciation of its beauty [51]. Image Aesthetic Assessment seeks to systematically appraise this aesthetic quality by analyzing the visual appeal of images [52]. Empirical psychological research corroborates that images can trigger a wide range of emotions, which are influenced by their aesthetic attributes and semantic content [53]. Previous research concentrated on aesthetic image captioning and analysis through the aggregation of commentary on aesthetic attributes [54]. These studies address the concepts of style, layout, and aesthetics from the viewpoints of beauty and visual attractiveness. Recent scholarly efforts have focused on encouraging vision-language models to generate visual connotations and captions related to various aesthetic attributes (e.g., color, harmony, lighting, composition) [55]. More recently, Kruk *et al.* [56] introduced a connotation-rich dataset, Impressions, designed to explore the emotions, thoughts, and beliefs that images evoke, along with the aesthetic elements that elicit these responses.

The introduction of this dataset marks a significant advance in the study of how visual stimuli can influence complex perceptual and emotional outcomes. In this study, we utilize aesthetic attributes to capture sentiment cues within visual content at both object and holistic levels. Inspired by Impressions [56], we further prompt the LLM to generate impression rationales for MASC, enabling analysis of the underlying affective resonance evoked by images.

2.3 LLMs-Based Rationale Generation

Recently, LLMs have achieved significant success across various downstream tasks [57]–[60]. LLMs such as GPT-4o [61], Gemini [62], and LLaMA-2 [63] hold significant potential to usher data annotation into a new era, functioning not merely as auxiliary tools but as vital enhancers of its effectiveness and quality [64], [65]. LLMs can automatically annotate samples, ensure consistency across large data volumes, and adapt to specific domains via fine-tuning, thereby establishing a new standard in deep learning [66]–[68]. The rationale represents the detailed cognitive process an individual typically follows when solving a problem, providing useful supplementary information for the final answer [69]. Early studies [70] typically relied on human experts to annotate rationale in datasets, significantly limiting availability and scalability. A bunch of diverse methodologies have been developed to produce high-quality and fine-grained rationale. Wang *et al.* [71] proposed to elucidate each choice in a sample by generating choice-specific rationales via LLMs. Wang *et al.* [72] enhanced the credibility of generated rationales by incorporating gold-standard answers and using contrastive decoding algorithms. Liu *et al.* [73] laid much emphasis on curating high-quality prompts to obtain fine-grained rationales from GPT-4o and build a logical chain-of-thought instruction-tuning dataset. More recently, Kang *et al.* [74] developed a sophisticated neural reranking mechanism to dynamically retrieve highly relevant supplementary documents for generating high-quality rationales in knowledge-intensive reasoning tasks.

In this paper, we build upon the work of Wang *et al.* [72] by fully utilizing the dataset's gold-standard annotations to generate semantic and impression rationales through meticulously designed prompts. This approach ensures high-quality rationale generation while avoiding additional costs from trial-and-error OpenAI API usage fees.

3 METHODOLOGY

This section presents our proposed framework for MASC, beginning with the task formalization, followed by the rationale dataset construction process, and concluding with the proposed method, comprising linguistic-aware semantic alignment, a translation module, rationale dataset construction and a rationale-aware learning framework.

3.1 Task Definition

Given a multimodal dataset M, each sample X_i consists of an image V_i paired with a sentence S_i containing one or more specific targets T_i. The goal of MASC is to predict the sentiment polarity $Y_i \in \{\text{Positive, Negative, Neutral}\}$ for a specific target T_i. Moreover, our framework infers both semantic rationale SR_i and impression rationale IR_i, explaining the sentiment prediction Y_i for a specific target T_i, based on multimodal semantic meaning and the affective resonance evoked by the image. In this study, the model outputs SR_i, IR_i, Y_i for an input sample $X_i = (S_i, V_i, T_i)$, where SR_i and IR_i offer supplementary sentimental cues for sentiment prediction T_i.

3.2 Method Overview

As shown in Figure 1, our proposed framework comprises four technical components, namely a Translation Module, Rationale Dataset Construction, Linguistic-aware Semantic Alignment, and Rationale-Aware Learning. The Translation Module converts visual content, both holistic and object-level, into language captions. For entire images, it generates emotion-laden aesthetic captions using our fine-tuned BLIP. For object-level content, it maps visuals to facial descriptions or aesthetic captions with rich emotional cues via EmoLA or our fine-tuned BLIP. The construction of the rationale dataset involves generating semantic and impression rationales. We curate prompts tailored to each rationale category and input them, along with the samples, into GPT-4o to collect the desired rationales. The Linguistic-aware Semantic Alignment module segments the input image into patches, dynamically selects and refines relevant visual patches, and achieves patch-token alignment guided by linguistic features from the input sentence. Lastly, we propose a Rationale-Aware Learning framework built up on a generative model that simultaneously learns sentiment classification, semantic rationale generation, and impression rationale generation from diverse textual inputs, such as sentences, aesthetic captions, and facial descriptions.

3.3 Translation Module

This module translates visual content into overall aesthetic captions, object-level facial descriptions, or object-level aesthetic captions in textual form, embedding rich sentimental cues to facilitate object-level sentiment alignment. Specifically, we leverage object annotations from the Fine-Grained Multimodal Named Entity Recognition (MNER) task [27], which annotates specific targets in the sentence and their corresponding objects in the image. The MNER dataset is derived from the same Twitter dataset as the MASC datasets, incorporating the original image-text pairs from MASC. We further pre-process the MNER dataset and transfer its object annotations to the MASC dataset. To generate aesthetic captions rich in sentimental cues, we fine-tune a BLIP model using the recent aesthetic-specific dataset, *Impression* [56]. For facial description, we deploy the LLM-based EmoLA [75] to interpret fine-grained human mental states from images.

To tackle the challenge of potential one-to-many annotation scenarios, wherein multiple visual objects correspond to a specific target in the sentence, we calculate the similarity between the entire image and all object annotations, retaining only the object with the highest similarity score for each specific target. Subsequently, we generate various textual auxiliary sentences, based on object annotations. Firstly, in cases where the object corresponding to a specific target is absent from the image, a fine-tuned BLIP model is applied to

generate an overall aesthetic caption $A^c = \left(a_1^c, a_2^c, \ldots, a_{N_c}^c\right)$ for the entire image:

$$A^c = BLIP_{fine}(V), \tag{1}$$

where $BLIP_{fine}(\cdot)$ is the fine-tuned BLIP over *Impression* dataset. If the object corresponding to a specific target is present in the image, we develop a Human-Object Differentiation (HOD) module based on the Sample and Computation Redistribution for Efficient Face Detection (SCRFD) [76] framework. This module determines the presence of a face within the annotated object-level visual content and assigns a facial binary label:

$$Y_i^{o_j} = HOD(V_i^{o_j}), \tag{2}$$

where $Y_i^{o_j} \in [1, 0]$ indicates whether the object-level visual content contains a face (0 for no face, 1 for face detected), and $V_i^{o_j}$ denotes the j-th object-level visual content in the i-th image. Subsequently, we generate facial descriptions or aesthetic captions for object-level visual content based on the facial binary label:

$$A^o = \begin{cases} EmoLA(V_i^{o_j}), & \text{if } Y_i^{o_j} = 1\ , \\ BLIP_{fine}(V_i^{o_j}), & \text{otherwise,} \end{cases} \tag{3}$$

where $A^o = \left(a_1^o, a_2^o, \ldots, a_{N_o}^o\right)$ is the generated auxiliary sentence (facial description or aesthetic caption) for the object-level visual content.

3.4 Rationale Dataset Construction

The current MASC benchmark includes only specific target (aspect) labels within the image-text pair sentences and their corresponding sentiment polarities. Recently, Fan *et al.* [26] introduced a dataset for MASC with cause analysis, focusing exclusively on textual semantics rather than integrating both visual and textual cues. Moreover, they overlook the affective resonance evoked by image aesthetic attributes, eliminating a crucial layer of emotional cues and resulting in an incomplete sentiment representation. This omission hinders the holistic integration of textual and visual modalities, leading to suboptimal sentiment modeling. Therefore, we employ GPT-4o to generate semantic and impression rationales, with the detailed generation process outlined in Algorithm 1.

Algorithm 1 Rationale Dataset Construction

Input: All samples (V, S, T, Y) in MASC dataset M
Output: Rationale dataset R which contains Semantic Rationale (SR) and Impression Rationale (IR)

```
Design & refine prompt pool for SR (SRP) and IR (IRP)
for each sample (V_i, S_i, T_i, Y_i) in M do
    //Randomly select a prompt from SRP for SR
    SR_prompt ← PromptPoolforSR(V_i, S_i, T_i, Y_i)
    //Randomly select a prompt from IRP for IR
    IR_prompt ← PromptPoolforIR(V_i, S_i, T_i, Y_i)
    Produce SR and IR via GPT-4o
    SR_i ← GPT-4o(V_i, S_i, T_i, Y_i, SR_prompt)
    IR_i ← GPT-4o(V_i, S_i, T_i, Y_i, IR_prompt)
    Add (V_i, S_i, T_i, Y_i, SR_i, IR_i) to R
end for
```

TABLE 1
Example prompts for semantic rationale generation.

Type	Prompts
System Prompt	You are an AI assistant specializing in multimodal understanding and sentiment analysis, particularly in scenarios involving the integration of image and text modalities.
Semantic Rationale Generation Prompt	You will be provided with an image-text pair. Your task is to analyze the sentiment towards the specified entity {aspect} and explain why the sentiment polarity {label} is appropriate. Your explanation should consider both the semantic meaning of the text and the visual representation of the image, focusing on explicit content and the emotional or contextual cues conveyed by their combination. Start your response with: "Based on the image-text pair, the sentiment towards {aspect} is {label} because...". Provide a concise, focused explanation highlighting the single most compelling reason for this sentiment classification.

To comprehensively capture the emotional rationale underlying the identified sentiment polarity from a semantic perspective of both image and text, we employ GPT-4o (gpt-4o-2024-05-13) via the OpenAI API[1] to generate SR. Meanwhile, to enable the model to effectively capture implicit emotional cues arising from the affective resonance of aesthetic attributes, we employ GPT-4o to generate the IR.

To enhance the diversity of generated semantic and impression rationales (SR and IR), we designed and refined a series of templates to construct separate prompt pools for SR and IR, from which a prompt is randomly selected as instructions to guide GPT-4o in generating the corresponding rationale. In this study, we adopt the approach outlined by Sarah *et al.* [77] and Wang *et al.* [72], leveraging tailored prompts conditioned on the dataset's gold-standard annotations to generate SR and IR using GPT-4o. The example prompts for generating SR and IR are presented in Tables 1 and 2, respectively.

3.5 Linguistic-aware Semantic Alignment(LSA)

We first introduce dynamic patch selection in Sec. 3.5.1. Then, we introduce the semantic patch calibration in Sec. 3.5.2. and patch-token alignment in Sec. 3.5.3. The overall process of LSA is shown in the persucode 2.

3.5.1 Dynamic Patch Selection(DPS)

Dynamic Patch Selection (DPS) is considered a discriminative task that assigns significance scores to visual patches and selects valuable patches based on high scores. For the image in an image-text pair, we opt for vision Transformers as the visual encoder. The image V is divided into N_v non-overlapping patches by spatial distribution. These patches are then input as a visual token sequence into the

1. https://platform.openai.com

TABLE 2
Example prompts for impression rationale generation.

Type	Prompts
System Prompt	You are an AI assistant specializing in multimodal emotion and aesthetic understanding, especially in analyzing the emotional responses elicited by visual content.
Impression Rationale Generation Prompt	You will be given an image-text pair. Your task is to analyze the specified entity {aspect} and its associated sentiment label {label} based entirely on the image's aesthetic attributes and the emotional resonance it conveys. Focus exclusively on the overall impression and visual connotations conveyed by the image, emphasizing why the assigned sentiment {label} aligns with the general mood or perception evoked by the entity. Avoid mentioning specific details; instead, highlight the prevailing emotional or aesthetic impression.

vision Transformer to obtain a set of visual patch features $V = (v_{cls}, v_1, v_2, \ldots, v_{N_v}) \in \mathbb{R}^{(N_v+1)\times d}$. For sentence S, a pre-trained Transformer serves as the textual encoder. The sentence is tokenized into N_s tokens and processed by the encoder to extract linguistic features $S = (s_1, s_2, \ldots, s_{N_s}) \in \mathbb{R}^{N_s\times d}$. Subsequently, we incorporate spatial information from images into visual patch features and use an MLP-based score-sensitive prediction mechanism to learn significant scores:

$$p_i^s = \mathit{Sigmoid}\left(\mathbf{MLP}\left(\boldsymbol{v}_i\right)\right), i \in \{1, 2, \ldots, N_v\}, \tag{4}$$

where $p_i^s \in [0, 1]$ represents the importance score assigned to each visual patch. Moreover, achieving refined cross-modal alignment requires more than depending solely on a scoring mechanism to identify valuable visual patches without linguistic supervision [78], [79]. Consequently, we introduce linguistic context by calculating attentive scores between visual patches and the input sentence. First, we derive linguistic-aware scores p_i^l through cross-attention between visual patches and linguistic features. Then, we enhance key visual content by computing self-attention within patches, producing image-prominent scores p_i^e:

$$p_i^l = Norm\left(\boldsymbol{v}_i \cdot S/d\right), p_i^e = Norm\left(\boldsymbol{v}_i \cdot V/d\right), \tag{5}$$

where $Norm(\cdot)$ denotes the normalization of scores to a range from 0 to 1. S and V represent the global embeddings for linguistic features and visual patches, respectively. These scores are integrated to derive the final value score:

$$p_i^f = (1-\beta)p_i^s + \frac{\beta}{2}\left(p_i^l + p_i^e\right), \tag{6}$$

where β refers to the weight parameter. After obtaining the value score $p^f = \left(p_1^f, p_2^f, p_3^f, \ldots, p_{N_v}^f\right) \in \mathbb{R}^{N_v}$, we convert it into a binary decision matrix $\{0, 1\}^{N_v}$ to determine patch selection. This matrix is constructed using the Gumbel-Softmax technique [80], ensuring a smooth and differentiable sampling process.

Algorithm 2 Linguistic-aware Semantic Alignment (LSA)

procedure DYNAMIC PATCH SELECTION(V, S)
  Extract visual patches $V \leftarrow \text{ViT}(V)$, text tokens $S \leftarrow \text{TextEnc}(S)$
  Compute significance scores: $p_i^s \leftarrow \text{MLP}(v_i)$, $p_i^l \leftarrow \text{Norm}(v_i S^\top)$, $p_i^e \leftarrow \text{Norm}(v_i V^\top)$
  Fuse scores: $p_i^f \leftarrow (1-\beta)p_i^s + \frac{\beta}{2}(p_i^l + p_i^e)$
  Apply Gumbel-Softmax sampling to obtain binary mask $D \in \{0,1\}^{N_v}$
  Return selected patches $V^p \leftarrow \{v_i | D_i = 1\}$
end procedure
procedure SEMANTIC PATCH CALIBRATION(V^p)
  Aggregate key patches: $\tilde{V}^p \leftarrow \text{Softmax}(\text{MLP}(V^p)) \cdot V^p$ ▷ Adaptive weighting
  Fuse redundant patches: $\tilde{v}^r \leftarrow \sum \tilde{p}_i v_i$ ▷ Weighted sum via p^f
  Return $\tilde{V}^p \leftarrow [v_{cls}; \tilde{V}^p; \tilde{v}^r]$
end procedure
procedure PATCH-TOKEN ALIGNMENT($\tilde{V}^p, S$)
  Compute cosine similarity matrix $A \in \mathbb{R}^{(N_f+2)\times N_s}$
  Calculate alignment score $K(V,S) \leftarrow \frac{1}{2}\left(\text{mean}(\max_j A_{ij}) + \text{mean}(\max_i A_{ij})\right)$
  Optimize with $\mathcal{L}_{\text{align}} \leftarrow$ Bi-directional Triplet Loss$(K(V,S), K(V,\hat{S}), K(\hat{V},S))$
end procedure

The Gumbel-Softmax matrix is defined as:

$$\boldsymbol{M}_{i,l} = \frac{\exp\left(\log\left(\boldsymbol{m}_{i,l} + G_{i,l}\right)/\tau\right)}{\sum_{j=1}^{L}\exp\left(\log\left(\boldsymbol{m}_{i,j} + G_{i,j}\right)/\tau\right)}, \tag{7}$$

where $M \in \mathbb{R}^{N_v\times L}$, L indicates the total number of categories. In this scenario, L is set to 2 for the binary decision ($\boldsymbol{m}_{i,1} = p_i^f$, $\boldsymbol{m}_{i,2} = 1 - p_i^f$). $G_i = -\log(-\log(U_i))$ represents the Gumbel distribution, U_i refers to the uniform distribution and τ is the temperature parameter.

Next, we obtain the differentiable decision matrix D by applying the arg-max on M:

$$\boldsymbol{D} = \text{Sampling}(\boldsymbol{M})_{*,1} \in \{0,1\}^{N_v}, \tag{8}$$

where D indicates patch selection outcomes: "1" for important patches and "0" for redundant ones. In the training stage, gradients are backpropagated through the differentiable decision matrix, enabling the dynamic selection of valuable visual patches via the score-sensitive prediction mechanism.

3.5.2 Semantic Patch Calibration(SPC)

This section aims to further refine the semantic representation of the selected valuable visual patches. After dynamically selecting important visual patches guided by linguistic supervision, we designate them as $V^p = \left(v_1^p, v_2^p, \ldots, v_{N_p}^p\right) \in \mathbb{R}^{N_p\times d}$. N_p is the number of selected valuable visual patches. We employ an aggregation network [81] to model multiple aggregation weights and combine the selected N_p visual patches to generate N_f informative visual features:

$$\tilde{\boldsymbol{v}}_j^p = \sum_{i=1}^{N_p}(\boldsymbol{W})_{ij} \cdot \boldsymbol{v}_i^p, \quad j = [1, \ldots, N_f], \tag{9}$$

$$\boldsymbol{W} = \textit{softmax}\left(\mathbf{MLP}\left(\boldsymbol{V}^{p}\right)\right), \tag{10}$$

where $(\boldsymbol{W})$ denotes the normalized weight matrix and $\sum_{i=1}^{N_s}(\boldsymbol{W})_{ij} = 1$. N_f is the number of aggregated patches ($N_f < N_p$). The aggregation network adaptively combines visually similar patches and is differentiable for end-to-end training. While redundant visual patches can be discarded, they may contain supplementary semantic features for refined cross-modal alignment. Therefore, we fuse them into a single patch:

$$\tilde{\boldsymbol{v}}^{r} = \sum_{i\in\mathcal{N}} \tilde{p}_i \cdot \boldsymbol{v}_i, \quad \tilde{p}_i = \frac{\exp\left(p_i^f\right)\boldsymbol{D}_i}{\sum_{i=1}^{N}\exp\left(p_i^f\right)\boldsymbol{D}_i}, \tag{11}$$

where $\mathcal{N}$ represents the set for redundant visual patches. $\tilde{p}_i$ denotes the normalized score of the value score p_i^f. Finally, this component models the calibrated refined visual patches, denoted as $\tilde{V}^p = \left(v_{cls}, \tilde{v}_1^p, \tilde{v}_2^p, \ldots, \tilde{v}_{N_f}^p, \tilde{v}^r\right) \in \mathbb{R}^{(N_f+2)\times d}$.

3.5.3 Patch-token Alignment(PTA)

This module aims to achieve the fine-grained patch-token level alignment. Specifically, we first utilize the refined visual patches $\tilde{V}^p$ and linguistic features S to compute token-wise similarities, producing a patch-token similarity matrix $A \in \mathbb{R}^{(N_f+2)\times N_s}$. $(A)_{ij} = \frac{(\tilde{\boldsymbol{v}}_i)^T \boldsymbol{s}_j}{\|\tilde{\boldsymbol{v}}_i\|\|\boldsymbol{s}_j\|}$ denotes the patch-token level alignment score between the i-th visual patch and the j-th word. Subsequently, maximum-correspondence interaction is introduced to aggregate cross-modal alignment. For each visual patch (or token), we identify the most aligned textual token (or patch) and calculate the average alignment score $K(V,S)$, representing the overall alignment between the image V and the sentence S:

$$K(V,S) = \frac{1}{N_f+2}\sum_{i=1}^{N_f+2}\max_j(\boldsymbol{A})_{ij} + \frac{1}{N_s}\sum_{j=1}^{N_s}\max_i(\boldsymbol{A})_{ij} \tag{12}$$

Following a previous method [82], the bi-direction triplet loss with hard negative mining is exploited:

$$\begin{aligned}\mathcal{L}_{\text{align}} = &\sum_{(V,S)}\left[\gamma - K(V,S) + K(V,\hat{S})\right]_+ \\ &+ \left[\gamma - K(V,S) + K(\hat{V},S)\right]_+,\end{aligned} \tag{13}$$

where γ is the trade-off parameter. $[x]_+ = \max(x,0)$ and (V,S) refers to a positive image-text pair in the mini-batch. Moreover, $\hat{S} = \text{argmax}_{j\neq\boldsymbol{S}}\, K(V,j)$ and $\hat{V} = \text{argmax}_{i\neq\boldsymbol{V}}\, K(i,V)$ indicate the hardest negative sentence and visual examples within a mini-batch, respectively.

3.6 Rationale-aware Learning

To endow the model with the ability to perform semantic causality and impression reasoning, we propose a rationale-aware learning framework designed to fine-tune a sequence-to-sequence (seq2seq) model. This seq2seq model is proposed to achieve three task objectives for each specific target within the image-text pair: sentiment classification (SC), semantic rationale generation (SRG), and impression rationale generation (IRG). These tasks are differentiated by the use of distinct input configurations and input content.

For SC, the decoder outputs only the predicted sentiment polarity. In SRG and IRG, the decoder produces the corresponding rationale and the sentiment prediction. Specifically, our input comprises the textual sentence $S = (s_1, s_2, \ldots, s_{N_s})$, the overall aesthetic caption of the image $A^c = \left(a_1^c, a_2^c, \ldots, a_{N_c}^c\right)$, the object-level description $A^o = \left(a_1^o, a_2^o, \ldots, a_{N_o}^o\right)$, which pertains to either facial or aesthetic attributes and the specific target T. The input format is determined by the presence of the specific target within the visual content. For example, if the specific target is identified in the image, based on the annotations provided by Wang *et al.* [27], the input for SC, SRG, and IRG is defined as follows:

$$H^{\text{sc}} = encoder(t_{\langle\text{sc}\rangle}, A^c, S, T), \tag{14}$$

$$H^{\text{srg}} = encoder(t_{\langle\text{srg}\rangle}, A^c, S, T), \tag{15}$$

$$H^{\text{irg}} = encoder(t_{\langle\text{irg}\rangle}, A^c, S, T), \tag{16}$$

where $encoder(\cdot)$ is the Transformer encoder of the seq2seq model. The tokens $t_{\langle\text{sc}\rangle}$, $t_{\langle\text{srg}\rangle}$, and $t_{\langle\text{irg}\rangle}$ are specialized tokens designed to represent distinct tasks. Although the specific aspects are not present in the image, this does not imply that sentimental cues from the image have no impact on predicting the sentiment polarity. On the contrary, incorporating sentiment cues from the holistic image can provide valuable insights into the influence of image aesthetic attributes on the sentiment prediction for the specific aspect. For samples where specific targets are present in the visual content, the input format is structured as follows:

$$H^{\text{sc}} = encoder(t_{\langle\text{sc}\rangle}, S, A^o, T), \tag{17}$$

$$H^{\text{srg}} = encoder(t_{\langle\text{srg}\rangle}, S, A^o, T), \tag{18}$$

$$H^{\text{irg}} = encoder(t_{\langle\text{irg}\rangle}, S, A^o, T). \tag{19}$$

We employ fine-grained, object-level emotion-laden descriptions to establish alignment between specific targets and their corresponding objects in the image, which enhances both the accuracy and interpretability of the sentiment prediction process. Subsequently, these hidden features are passed through a stack of self-attention-based encoders, which dynamically fuse representations and model both intra-modal and cross-modal interactions. Finally, the decoder produces task-specific outputs. For Sentiment Classification (SC), the decoder generates the predicted sentiment polarity, selecting from "positive," "negative," or "neutral," denoted as $\hat{y}_{sc}$:

$$G^{\text{sc}} = \left[\langle\text{sen}\rangle\hat{y}^{\text{sc}}\langle/\text{sen}\rangle\right], \tag{20}$$

where the special tokens $\langle\text{sen}\rangle$ and $\langle/\text{sen}\rangle$ are denoted as the start and end markers for SC predictors. For the two additional rationale generation tasks SRG and IRG, the decoder generates not only the semantic rationale $\hat{sr}$ and impression rationale $\hat{ir}$ for the specific target but also their corresponding sentiment predictions $\hat{y}_{sr}$ and $\hat{y}_{si}$:

$$G^{\text{sr}} = \left[\langle\text{sr}\rangle\hat{sr}\langle/\text{sr}\rangle\langle\text{sen}\rangle\hat{y}^{\text{sr}}\langle/\text{sen}\rangle\right], \tag{21}$$

$$G^{\text{ir}} = \left[\langle\text{ir}\rangle\hat{ir}\langle/\text{ir}\rangle\langle\text{sen}\rangle\hat{y}^{\text{ir}}\langle/\text{sen}\rangle\right], \tag{22}$$

where $\langle\text{sr}\rangle$, $\langle/\text{sr}\rangle$, $\langle\text{ir}\rangle$, $\langle/\text{ir}\rangle$, $\langle\text{sen}\rangle$, and $\langle/\text{sen}\rangle$ serve as

specialized markers to delineate the rationale and sentiment polarity. Finally, the input sequence is uniformly denoted as X, and the generated textual content is represented as $Z = \{z_1, z_2, \ldots, z_{N_z}\}$. Consequently, the loss function for the generation process is formulated as follows:

$$\mathcal{L}_Z = -\frac{1}{N}\sum_{i=1}^{N}\sum_{n_z=1}^{N_z}\log P\left(z_{i,n_z} \mid \hat{z}_{i,<n_z}, X\right), \quad (23)$$

where z_{i,n_z} is the ground truth token at position n_z for sample i, $\hat{z}_{i,<n_z}$ represents the generated sequence up to position n_z-1 for sample i, and $P(z_{i,n_z} \mid \hat{z}_{i,<n_z}, X)$ denotes the probability of generating token z_{i,n_z} conditioned on $\hat{z}_{i,<n_z}$ and X. In this rationale-aware learning framework, since all objectives are formulated as generative tasks, the loss functions $\mathcal{L}_{SC}$, $\mathcal{L}_{SRG}$, and $\mathcal{L}_{IRG}$ are all employ the generative loss function, E.q. 23. Therefore, the objective function in the proposed method is formulated as follows:

$$\mathcal{L} = \alpha\mathcal{L}_{\mathrm{SC}} + \frac{1-\alpha}{2}\mathcal{L}_{\mathrm{SRG}} + \frac{1-\alpha}{2}\mathcal{L}_{\mathrm{IRG}} + \lambda\mathcal{L}_{align}, \quad (24)$$

where $\alpha, \lambda \in (0, 1)$ are tradeoff hyperparameters that regulate the relative contributions of each generative loss and the patch-token alignment.

4 EXPERIMENTS

In this section, we provide a comprehensive description of the experimental settings and evaluate the proposed method on three publicly available MASC datasets, benchmarking it against state-of-the-art methods. Furthermore, we perform an extensive series of studies to thoroughly analyze the effectiveness of the proposed approach.

4.1 Experimental Settings

4.1.1 Datasets

We utilize three widely recognized benchmark datasets for MASC [13], [83]: Twitter-2015, Twitter-2017, and the Political Twitter dataset. Each sample within these datasets comprises a user-generated multimodal image-text pair, including an image, a textual sentence, and one or more specific targets. Each aspect is annotated with a sentiment label from the set Positive, Negative, Neutral. The detailed statistics of these datasets are presented in Table 3. Furthermore, we incorporate semantic rationale (SR), impression rationale (IR), aesthetic captions for the entire image (AC), facial descriptions (FD), and aesthetic captions for objects (AO) for each data point. The maximum length for facial descriptions and aesthetic captions is constrained to 50 tokens.

4.1.2 Implementation Details

We adopt the seq2seq model Flan-T5 [84] as the backbone of our generative framework. Specifically, the model is trained for 10 epochs using the AdamW optimizer [85], with a batch size of 4. A grid search is performed on the development set to determine the optimal learning rate, α and λ for Flan-T5 across the three datasets. The selected values for learning rate are $3e-4$, $3e-4$, $1e-4$, respectively, for the Twitter-2015, Twitter-2017 and Political Twitter. The trade-off hyperparameter sets (α and λ) are 0.2, 0.1, 0.2 and 0.2, 0.5, 0.5, respectively, for the Twitter-2015, Twitter-2017 and Political Twitter. Consistent with prior research on MASC [13], [26], we employ Accuracy (Acc) and F1 score (F1) as the evaluation metrics. The model is implemented using PyTorch, and experiments are conducted on an NVIDIA V100 GPU with 30 GB of memory.

4.2 Compared Baselines

We conducted a comprehensive comparative evaluation of the proposed method against a range of robust baseline approaches, which are classified into three categories. The first category consists of image-only methods:

- **Res-Target** [86] leverages ResNet as its backbone to extract visual features exclusively for predicting the sentiment of the specified target.

The second category includes text-only approaches:

- **MemNet** [87] employs a stacked architecture of multiple memory networks to build deep memory networks.
- **MGAN** [88] is based on a multi-grained attention architecture designed to adaptively capture both coarse-grained and fine-grained interactions.
- **BERT** [89] is a powerful pre-trained language model trained using a masked language modeling objective and next sentence prediction.

Finally, this study incorporates the following advanced image-text multimodal approaches:

- **MIMN** [90] comprises two customized interactive memory networks designed to capture inter-modal dynamics between different modalities and intra-modal dynamics within each individual modality.
- **ESAFN** [14] is a target-sensitive interaction and fusion network designed to adaptively capture interactive features across modalities while also modeling intra-modality features.
- **TomBERT** [13] utilizes BERT and ResNet as backbone models for encoding textual and visual content, respectively. Cross-modal fusion is accomplished by integrating these features into a BERT encoder.
- **JML-MASC** [46] jointly extracts the specific targets and identifies their sentiment polarity by utilizing a visual de-nosing mechanism and attention-based fusion framework.
- **EF-CapTrBERT** [19] converts visual content into an auxiliary sentence, which is then combined with the input sentence and processed through a BERT encoder for sentiment prediction.
- **VLP-MABSA** [16] is a task-specific pre-trained generative framework for multimodal aspect-based sentiment analysis, built on the BART architecture.
- **FITE** [25] is a translation-based approach, which captures facial features in the image and translates them into a corresponding facial description as an auxiliary sentence for sentiment classification.
- **CMMT-MASC** [17] is a cross-modal multi-task Transformer designed for MASC. Additionally, it employs multimodal gating mechanisms to dynamically regulate the flow of textual and visual information during interactions.

TABLE 3
Detailed Statistics of Twitter-2015, Twitter-2017, and Political Twitter datasets. The "#sentence" refers to the total number of sentences. "#Avg. Length" denotes the average length of sentences, while "#Avg. Aspect" indicates the average number of aspects in a sentence. "#Avg. Length of SR", "#Avg. Length of IR", "#Avg. Length of AC", "#Avg. Length of FD", and "#Avg. Length of AO" correspond to the average lengths of semantic rationales (SR), impression rationales (IR), aesthetic captions for the entire image, facial descriptions, and aesthetic captions for objects.

Label	Twitter-2015			Twitter-2017			Political Twitter		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
Positive	928	303	317	1508	515	493	3318	570	176
Neutral	1883	670	607	1638	517	573	4697	823	368
Negative	368	149	113	416	144	168	887	166	305
Total	3179	1122	1037	3562	1176	1234	8902	1559	849
#Sentence	2101	727	674	1746	577	587	5105	900	407
#Avg. Length	16.72	16.74	17.05	16.21	16.37	16.38	16.62	16.67	16.59
#Avg. Aspect	1.51	1.54	1.54	2.04	2.04	2.10	1.74	1.73	2.09
#Avg. Length of SR	42.5	42.4	42.5	42.6	42.8	43.0	42.7	42.6	42.2
#Avg. Length of IR	56.7	56.0	55.7	55.5	56.1	55.4	55.9	56.1	56.3
#Avg. Length of AC	35.9	35.9	35.5	32.5	32.5	31.6	34.0	34.2	33.3
#Avg. Length of FD	39.2	38.5	37.8	38.9	38.5	39.3	39.0	38.4	38.7
#Avg. Length of AO	29.1	29.7	30.3	28.9	29.4	28.9	29.1	29.1	31.3

- **HIMT** [91] is a Transformer framework that incorporates a hierarchical interaction component to model the relationships between specific aspects and the input sentence.
- **IMT** [15] is a coarse-to-fine-grained multimodal matching network that predicts image-target relevance and performs object-target alignment to support sentiment polarity identification.
- **CoolNet** [21] is a fine-grained cross-modal alignment approach that aligns textual and visual content from both semantic and syntactic perspectives.
- **UnifiedTMSC** [92] introduces a descriptive prompt paraphrasing paradigm to generate paraphrased prompts, while optimizing image vectors within the multimodal space of vision and language.
- **VEMP** [93] decodes the semantic of visual elements by utilizing textual tokens in the image, target-aware adjective-noun pairs, and image captions.
- **Atlantis-MASC** [24] is a trident-shaped, aesthetics-driven approach for joint MABSA, which integrates image aesthetic and achieves effective alignment of vision and text across multiple granular levels.
- **MDCA** [26] is a generative framework proposed to provide explicit rationales to explain why specific content conveys certain sentiment.

4.3 Main Results

The main results are presented in Table 4. Given that the two additional rationale generation tasks contribute to improving sentiment prediction by providing explanations for the underlying causes of sentiment, we select the prediction results from sentiment classification $\hat{y}^{\text{sc}}$ as the primary outcomes for accuracy and F1 score evaluation. As presented in Table 4, the proposed method demonstrates competitive performance on both Twitter datasets compared to strong baselines from both text-only and multimodal approaches.

Specifically, it achieves the highest accuracy (81.61%) and F1 score (77.98%) on the Twitter-2015 dataset, as well as the best accuracy (75.62%) and a near-optimal F1 score (74.59%) on the Twitter-2017 dataset. Compared to the image-only approach (Res-Target), the proposed method achieves a remarkable improvement of over 21.73% in accuracy on the Twitter-2015 dataset. Similarly, when compared to the best-performing text-only method (BERT), the proposed method demonstrates a substantial performance gain, with a 7.46% increase in accuracy and a 9.12% improvement in F1 on Twitter-2015. These observations underscore the limitations of single-modality approaches in capturing subtle sentiment cues from multimodal content. Moreover, the proposed method consistently outperforms recent multimodal models, such as UnifiedTMSC, Atlantis-MASC, and MDCA. For instance, UnifiedTMSC adopts a paraphrasing-based approach to enrich textual features but lacks explicit modeling of visual aesthetic-driven affective impact. On Twitter-2017, the proposed method achieves comparable F1 performance (74.59 vs. 74.70) while delivering higher accuracy (75.62 vs. 75.40), which highlights the complementary benefits of aesthetic affective resonance modeling. Although Atlantis-MASC incorporates image aesthetics, it mainly relies on global alignment techniques, which may overlook the intricate relationships between aspects and objects. The proposed method surpasses Atlantis-MASC by 1.58% in accuracy on Twitter-2017, underscoring the efficacy of its patch-token level and object-level alignment in capturing aspect-specific visual details. While MDCA incorporates reasoning and direct causality to explain sentiment causes, it primarily emphasizes textual semantic reasoning, which restricts its ability to capture visual content and the corresponding aesthetic affective resonance. In contrast, the proposed method surpasses MDCA with a 0.90% improvement in accuracy and a 0.83% increase in F1 on the Twitter-2015 dataset. This gain shows the benefits of jointly modeling semantic and affective resonance in sentiment causality.

TABLE 4
The main results (%) are presented with the best-performing results highlighted in **bold** and the second-best values indicated with underlined text.

Modality	Model	Venue	Twitter-2015		Twitter-2017		Political Twitter	
			Acc	F1	Acc	F1	Acc	F1
Image Only	Res-Target	CVPR 2016	59.88	46.48	58.59	53.98	60.21	58.42
Text Only	MemNet	EMNLP 2016	70.11	61.76	64.18	60.90	-	-
	MGAN	EMNLP 2018	71.17	64.21	64.75	61.46	67.37	62.78
	BERT	NAACL 2019	74.15	68.86	68.15	65.23	69.41	64.25
Image and Text	MIMN	AAAI 2019	71.84	65.69	65.88	62.99	70.52	65.39
	ESAFN	TASLP 2019	73.38	67.37	67.83	64.22	69.22	64.66
	TomBERT	IJCAI 2019	77.15	71.15	70.34	68.03	69.65	62.35
	JML-MASC	EMNLP 2021	78.70	-	72.70	-	70.14	68.37
	EF-CapTrBERT	ACM MM 2021	78.01	73.25	69.77	68.42	69.04	64.94
	VLP-MABSA	ACL 2022	78.60	73.80	73.80	71.80	70.32	69.64
	CMMT-MASC	IPM 2022	77.90	-	73.8	-	-	-
	FITE	EMNLP 2022	78.49	73.90	70.90	68.70	68.64	65.83
	HIMT	TAFFC 2022	78.14	73.68	71.14	69.16	-	-
	IMT	IJCAI 2022	78.27	74.19	72.61	71.97	69.92	67.86
	CoolNet	IPM 2023	79.92	75.28	71.64	69.58	70.91	70.25
	UnifiedTMSC	EMNLP 2023	79.80	76.30	75.40	**74.70**	-	-
	VEMP	EMNLP 2023	78.88	75.09	73.01	72.42	-	-
	Atlantis-MASC	INFFUS 2024	79.03	-	74.20	-	69.83	68.97
	MDCA	TNNLS 2024	80.71	77.15	73.91	72.37	71.38	70.94
Ours	Chimera	TAFFC 2025	**81.61**	**77.98**	**75.62**	74.59	**72.56**	**72.32**

4.4 Results on Political Twitter

The Political Twitter dataset differs significantly from Twitter-2015 and Twitter-2017, especially due to its challenging domain shift between training, development, and test sets. Such domain differences create substantial barriers to generalization, which makes the task particularly suitable for advanced models that can comprehend subtle causality and context shifts.

From Table 4, it can be observed that the proposed Chimera demonstrates distinct advantages over existing approaches on the Political Twitter dataset. Compared to the third best performing method CoolNet, which achieved 71.32% accuracy and 69.64% F1 score, Chimera showcases a significant improvement. Similarly, MDCA, which performed with an accuracy of 71.38% and an F1 score of 70.94%, still lags behind Chimera. Additionally, we observed that the discrepancy between accuracy and F1-score significantly narrows as accuracy increases, particularly when accuracy surpasses 70%.

We hypothesize that the underlying cause may lie in the relatively balanced class distribution of sentiment categories (e.g., positive, neutral, negative) within the Political Twitter test set (as shown in Table 3). At higher accuracy levels, the ratios of false positives to false negatives exhibit increasing symmetry across models. This equilibrium consequently reduces the divergence between precision and recall metrics, thereby causing the F1-score – defined as their harmonic mean – to naturally converge with accuracy.

4.5 Ablation Study

To systematically investigate the influence of the linguistic-aware semantic alignment module, including semantic and impression rationale reasoning as well as object-level fine-grained alignment, on sentiment prediction, we conducted ablation studies and the results are shown in Table 5. As presented in Table 5, the exclusion of semantic rationale ("w/o SRG") results in a noticeable performance decline across all three datasets. This effect is particularly pronounced on the Twitter-2017 and Political Twitter datasets, where nearly all evaluation metrics, including accuracy and F1 score, exhibit a reduction of approximately 2%. Similarly, the absence of impression rationale reasoning ("w/o IRG") results in performance fluctuations on the Twitter-2015 and Political Twitter datasets. However, the most noticeable effect is observed on the Twitter-2017 dataset, where the model's performance exhibits a significant degradation, particularly in the sentiment classification task, with nearly a 4% drop in both accuracy and F1 score. The results ("w/o IRG & AC") reveal consistent performance degradation in both Accuracy and F1-score across all three datasets. Particularly noteworthy is the model's inferior performance on Twitter-2017 and Political Twitter datasets compared to the baseline(w/o IRG). However, an unexpected performance improvement emerges in Twitter-2015, surpassing even the configuration retaining aesthetic captions as input. This phenomenon may be attributed to dataset-specific characteristics in sample distribution.

TABLE 5
The results (%) of the ablation study for our Chimera model are presented. The top-performing values emphasized in **bold** and the second-best values distinguished using underlined text. The notations "w/o SRG," "w/o IRG," and "w/o SRG & IRG" denote the exclusion of the respective generative tasks. "w/o IRG & AC" refers to the removal of IR generation task and replace the aesthetic caption (AC) with general caption. "w/o LSA" represents the removal of the Linguistic-aware Semantic Alignment branch, while "w/o OD" indicates the exclusion of object-level descriptions (e.g., facial descriptions and object-level aesthetic captions) from the input sequence.

Method	Twitter-2015						Twitter-2017						Political Twitter					
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
	SC		SRG		IRG		SC		SRG		IRG		SC		SRG		IRG	
Chimera	**81.61**	**77.98**	81.12	77.11	77.56	73.55	**75.62**	**74.59**	75.09	73.64	71.96	68.23	**72.56**	**72.32**	71.69	71.40	69.30	68.95
w/o SRG	80.52	76.10	-	-	75.83	70.96	73.50	72.49	-	-	70.66	67.20	70.43	69.88	-	-	68.25	67.58
w/o IRG	80.23	75.22	80.03	75.42	-	-	71.88	70.16	72.6	70.73	-	-	71.15	70.70	71.01	70.52	-	-
w/o IRG & AC	80.67	76.03	80.11	76.46	-	-	71.59	69.83	72.25	70.33	-	-	70.62	70.06	71.04	70.47	-	-
w/o SRG & IRG	77.24	71.82	-	-	-	-	71.23	68.98	-	-	-	-	67.88	67.20	-	-	-	-
w/o LSA	80.54	77.03	79.75	76.22	76.52	72.03	73.72	70.96	74.38	72.26	71.36	67.88	71.86	71.37	70.92	70.55	68.43	67.99
w/o OD	79.96	76.08	80.09	76.32	77.12	72.84	73.06	70.85	74.37	72.36	71.11	67.53	71.64	71.12	71.12	70.77	68.55	68.07
w/o Aes-cap	80.03	75.27	79.94	76.05	75.69	71.08	72.36	71.64	72.28	71.21	69.28	65.44	69.43	68.94	69.37	69.00	67.85	67.27

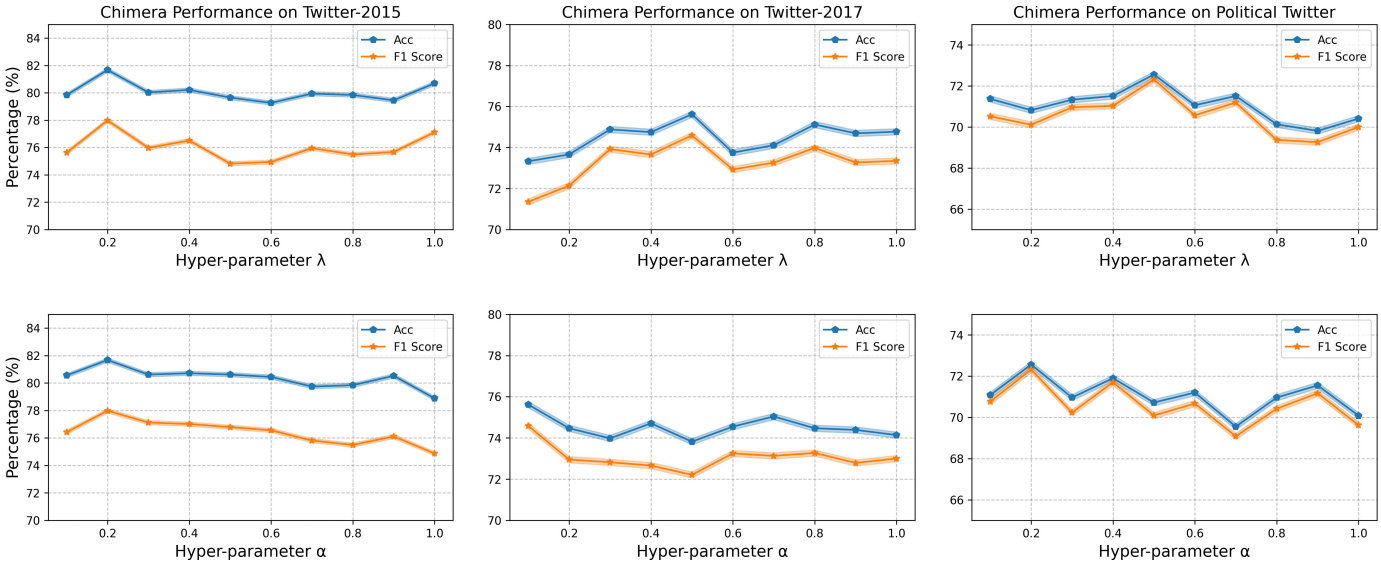

Fig. 2. Results (%) on hyper-parameter of α and λ.

As detailed in Table 3, Twitter-2015 exhibits a significantly higher proportion of neutral-class samples compared to Twitter-2017 and Political Twitter. When the Chimera model is deprived of its reasoning abilities for both semantic and impression rationales ("w/o SRG & IRG"), its performance on sentiment classification declines to the lowest levels across all datasets. Specifically, a consistent reduction of approximately 4-5% is observed in nearly all metrics, underscoring the essential role of rationale-based reasoning in enhancing the effectiveness and accuracy of sentiment analysis tasks. These results show that the influence of rationale reasoning differs across datasets. For Twitter-2017, with its balanced sentiment distribution (see Table 3), impression rationale has a greater impact on sentiment analysis. In contrast, both semantic and impression rationales contribute to the other two datasets, but neither is dominant.

The LSA branch plays a pivotal role in the Chimera model by bridging the semantic gap between textual and visual modalities, ensuring effective alignment of information across visual and textual data. Its removal (w/o LSA) consistently leads to a significant decline in performance across all datasets, as evident in the ablation study. For instance, on Twitter-2015, the accuracy drops from 81.61% to 80.54%, and the F1 score decreases from 77.98% to 77.03%. Similarly, for Twitter-2017, accuracy, and F1 score dropped to 73.72% and 70.96%, respectively. By aligning linguistic and visual features, the branch allows the model to effectively interpret semantic overlaps and contrasts, enabling more accurate sentiment predictions.

Object-level descriptions (e.g., facial expressions and object-level aesthetic captions) enrich the input sequence by providing object-level detailed visual context. The ablation study reveals that removing OD (w/o OD) causes noticeable performance drops. On Twitter-2015, accuracy drops by 1.65 percentage points, and the F1 score decreases by 1.90 percentage points. Similarly, on Twitter-2017, accuracy is reduced by 2.56 percentage points, while the F1 score drops by 3.74 percentage points. Without the OD, the model loses access to these fine-grained visual features, leading to diminished interpretability and accuracy, particularly in datasets where visual information plays a crucial role in determining sentiment. Additionally, the aesthetic caption is excluded from the input sequence to assess its impact on performance (w/o Aes-cap). As demonstrated in Table 5, the absence of aesthetic features results in a noteworthy decline in performance across all datasets, particularly in

TABLE 6
The evaluation resutls (%) of rationale quality. The best-performing results highlighted in **bold**.

Rationale Source	Twitter-2015		Twitter-2017		Political	
	Acc	F1	Acc	F1	Acc	F1
	Semantic Rationale					
Ground-Truth	**99.04**	**99.04**	**98.54**	**98.54**	**97.64**	**97.64**
Chimera	80.91	80.83	75.04	74.93	70.20	70.14
	Impression Rationale					
Ground-Truth	**69.91**	**69.90**	**72.77**	**72.71**	**76.8**	**76.87**
Chimera	63.45	63.65	61.67	59.38	60.54	60.12

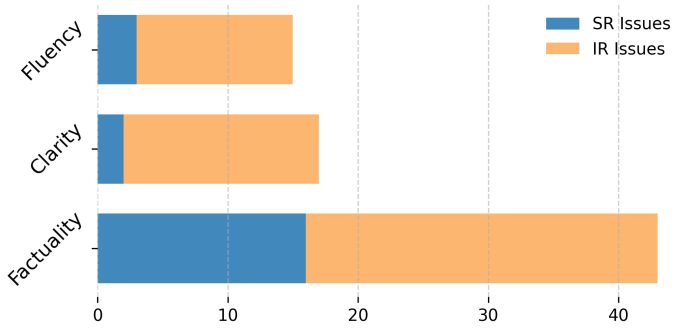

Fig. 3. Human evaluation of factuality, clarity and fluency for SR and IR.

the impression rationale generation (IRG) task. This leads to Chimera exhibiting the poorest sentiment classification performance for IRG on the Twitter-2017 and Political Twitter datasets, which underscore the importance of aesthetic captions in guiding the model to generate coherent and emotionally nuanced impressions.

4.6 Hyper-parameter Analysis

We conduct a hyperparameter analysis to explore the impact of α and λ on the Chimera model's performance across the Twitter-2015, Twitter-2017, and Political Twitter datasets. Hyperparameter α regulates the balance between sentiment classification (SC) and rationale generation components (semantic and impression rationales, SRG, and IRG), while λ controls the weight of patch-token alignment within the overall loss function. As shown in Figure 2, for all datasets, a lower α, which assigns greater weight to rationale generation, generally improves model performance, with values around 0.1 to 0.2 achieving the highest accuracy and F1 scores. This emphasizes the significance of integrating semantic and impression rationales in MASC. As α increases, favoring SC loss, performance plateaus or declines, particularly for the Political Twitter dataset, indicating that reduced emphasis on rationale generation diminishes the model's ability to capture fine-grained sentiment context effectively. Moreover, the results indicate that increasing λ initially enhances model performance, with diminishing returns beyond a certain threshold. For the Twitter-2015 and Political Twitter datasets, moderate λ values [0.2, 0.5] achieve optimal accuracy and F1 scores, while higher values ($\lambda > 0.6$) lead to performance stabilization or slight decline. This observation indicates that balanced alignment between visual and textual features enhances the model's interpretability and accuracy and excessively high λ values may negatively impact performance, likely due to overemphasis on alignment at the expense of core sentiment classification. For Twitter-2017, a similar trend is observed, although performance variations are less pronounced.

5 In-depth Analysis

5.1 Quality Analysis of Rationale

Table 6 provides an evaluation of the sentiment rationale quality for both the ground-truth and Chimera-generated content, aiming to analyze their impact on sentiment analysis. A pre-trained sentiment classification model [94] is employed to assess the intuitive sentiment quality of these rationales across three test datasets by inputting the rationales into the model and analyzing the sentiment predictions. For both SR and IR, the results in the Ground-Truth row represent the upper performance bound. It is evident that the ground truth performance for SR significantly exceeds that of IR, indicating that semantic rationales are more critical for this task than impression rationales. We hypothesize that two factors contribute to this discrepancy. Firstly, as illustrated in Table 3, semantic rationales are shorter in length and straightforward, facilitating easy comprehension, while the emotions elicited by images are inherently more abstract and multifaceted. Secondly, the IR's reliance on visual cues contrasts sharply with the Twitter dataset's text-centric sentiment distribution. Prior research has shown that a considerable majority of targets (around 58%) are absent from images [15], and most targets (93% in Twitter-2015) exhibit emotional coherence with their textual counterparts [95]. This misalignment underscores the dataset's limitations in evaluating IRs and necessitates a nuanced understanding of the interplay between visual and textual sentiment representations.

A total of 180 samples were randomly selected for human evaluation, with 100 samples drawn from the training set, 40 from the testing set, and 40 from the validation set of both the Twitter-2015 and Twitter-2017 datasets. Four native English speakers with Master's degrees in the arts were recruited to assess the quality of the rationale data based on three criteria: (1) factuality, evaluating whether the rationale is grounded in accurate and verifiable information; (2) clarity, assessing the logical structure and comprehensibility of the rationale; and (3) fluency, measuring the grammatical accuracy and smoothness of the language used. The Fleiss' Kappa (κ) values for the initial evaluation across the four raters were as follows: factuality $\kappa = 0.922$, clarity $\kappa = 0.945$, and fluency $\kappa = 0.960$. In cases of disagreement, the evaluators engaged in discussions to reach a consensus. Figure 3 presents the results of the human evaluation. It can be observed that SR consistently exhibits higher quality across all metrics, which verifies that the employed LLM is capable of generating appropriate rationale data for specific tasks when provided with concrete ground-truth labels. In comparison to SR, IR demands a more in-depth understanding of visual content and is inherently more subjective.

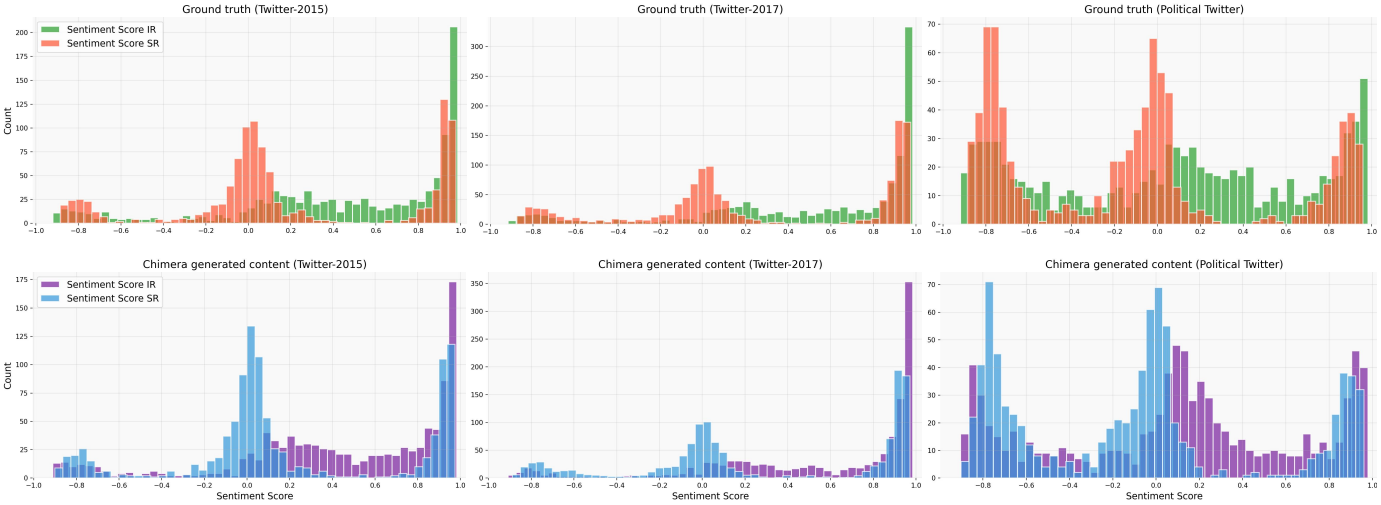

Fig. 4. Assessment of sentiment intensity for SR and IR in both ground truth data and Chimera-generated content.

Consequently, IR is more prone to issues of factuality and clarity, as interpreting the abstract aesthetic and emotional elements conveyed by an image often involves subjective reasoning, which may lead to misalignment with objective ground truths or human expectations.

5.2 Quantitative Analysis of Rationale

We conduct a quantitative analysis on the test sets of ground truth and Chimera-generated content to examine the impact of varying levels of sentiment intensity in cognitive rationales on the accuracy of sentiment prediction, including their potential to amplify or diminish predictive performance. As illustrated in Figure 4, the sentiment intensity distributions of Twitter-2015 and Twitter-2017 reveal distinct patterns. Specifically, the sentiment intensity of IR demonstrates a noticeable bias toward positive values, whereas the sentiment intensity of SR aligns more closely with the sentiment polarity label distribution presented in Table 3.

This observation suggests that IR demonstrates a bias toward positive samples, increasing the model's confidence in predicting positive instances. While this bias may be beneficial for datasets with a higher proportion of positive samples (e.g., Twitter-2017), it could lead to additional bias in datasets with a limited representation of positive samples. This finding is further corroborated by the ablation study results, which reveal that the performance of the Chimera model without IR is worse on Twitter-2017 compared to its performance on Twitter-2015. Another notable observation is that, for the ground truth of the Political Twitter dataset, the sentiment intensity distribution of IR is relatively uniform across all ranges. In contrast, the Chimera-generated content for IR exhibits a more distinguishable sentiment intensity distribution compared to the ground truth, which further validates the quality of SR, the effectiveness of the proposed Chimera training paradigm, and the robustness of Chimera's performance.

5.3 Impact of Aesthetic Attributes on Sentiment

To investigate the impact of image aesthetic attributes on sentiment analysis, we visualize the frequency of aesthetic-related words within the impression rationales generated

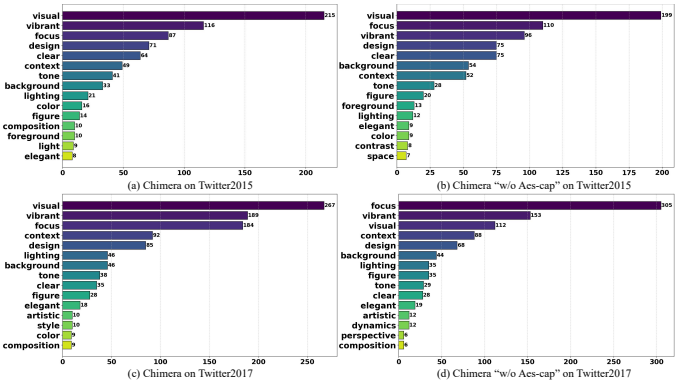

Fig. 5. Visualization of the top 15 most frequent aesthetic-related words in generated IR.

by our proposed Chimera model and its variant "Chimera w/o Aes-cap" on the Twitter-2015 and Twitter-2017 test sets. Specifically, we visualize the top 15 most frequent aesthetic-related words within the generated IR, based on the aesthetic attributes defined by Milena *et al.* [96]. As shown in Figure 5, the frequency analysis of aesthetic-related words for Chimera on Twitter-2015 and Twitter-2017 reveals that "visual," "vibrant," "focus," and "design" prominently appear across both datasets. These terms, associated with visual clarity, expressive quality, image composition, and cohesiveness, align with the model's improved accuracy and F1 scores. However, excluding the aesthetic caption from the input results in subtle shifts in the frequency distribution of these aesthetic-related terms. For Twitter-2015, the overall frequency distribution of aesthetic-related terms shows minimal change, with a slight increase in "focus" and a decrease in "vibrant". In contrast, for Twitter-2017, "focus" shows a significant increase, while "visual" and "vibrant" decrease notably. Combined with the ablation study results in Table 5, the absence of aesthetic captions in the input leads to the worst sentiment analysis performance across all datasets on IRG. This highlights the critical role of aesthetic captions in enhancing the model's understanding of image aesthetics, particularly in datasets like Twitter-2017 with bal-

Input image-text pair & Auxiliary sentences	Method	Semantic Rationale (SR of GT, Chimera and ChatGPT-4o)/ Reasoning Cause(RC of MDCA)	Impression Rationale (IR of GT and Chimera)/ Direct Cause(DC of MDCA)	Sentiment
RT @ MiamiHerald : [Joanne Stiger]$_{POS}$ and Melissa Keller are the first same - sex couple to be married in Broward :  Aesthetic caption: The body language of the two people shows that they are in a very different way. the background shows a room with a lot of people and the cameraman is in the background.	Ground Truth (GT)	Based on the image-text pair, the sentiment towards Joanne Stiger in the given text and image is positive, because the text highlights a historic and celebratory event of her marriage, and the image shows people smiling and appearing joyful, indicating happiness and approval.	The image emanates a joyous and celebratory atmosphere, highlighted by the smiling faces and relaxed postures of the individuals present. The presence of photographers and the gathering of people around the couple enhance the sense of an important and positive milestone being reached. Overall, the visual conveys happiness, love, and a significant moment being warmly embraced by all involved.	Positive 😀
	Chimera	The sentiment towards Joanne Stiger in the given text and image is positive, because the text highlights her as the first same-sex couple to be married in Broward, and the image shows her smiling and looking happy, which conveys a sense of pride and joy.	The image depicts two individuals who appear relaxed and happy, with one person appearing content and relaxed. The overall impression is one of happiness and contentment. The setting and body language suggest a sense of happiness, together with a pleasant and celebratory atmosphere.	Positive 😀 ✓
	MDCA	Because the tweet simply reports on a factual event without expressing any emotional response towards Joanne Stiger. The use of the word "first" suggests that the author is excited about the news and is highlighting the fact that Stiger and Keller are the first same-sex couple to be married in Broward. Overall, the tweet does not provide enough information to determine a positive or negative sentiment towards JoAnn Stiger herself.	Because joanne stiger and Melissa keller are the first same - sex couple to be married in bronward.	Neutral 😐 ✗
	GPT-4o	Based on the image-text pair, the sentiment towards Joanne Stiger in the given text and image is Positive, because both the faces of Joanne Stiger and Melissa Keller, as well as the individuals around them, display expressions of happiness and joy, indicating a celebratory and momentous occasion.		Positive 😀 ✓
Primary Football Cup . [St Bede]$_{NEU}$ ' s v St Augustines friendly . . . . both out but now for the friendly # sportsmanship  Aesthetic caption: The position of the camera captures the players in action, as they are engaged in a game of soccer. the position of the camera captures the players in action, and the position of the camera captures the players in action.	Ground Truth	Based on the image-text pair, the sentiment towards St Bede in the given text and image is neutral, because the text describes the event as a friendly match with a focus on sportsmanship, and the image shows a regular football game without any emotional or evaluative expressions towards St Bede.	The image depicts a friendly football match between St Bede's and St Augustine's on a lush green field under a partly cloudy sky. The overall scene conveys a sense of camaraderie and sportsmanship, aligning well with the neutral sentiment label. The focus appears to be on enjoying the game and fostering friendly competition in a relaxed atmosphere.	Neutral 😐
	Chimera	The sentiment towards St Bede in the given text and image is neutral, because the text simply states that St Bedé is participating in a friendly match without expressing any positive or negative emotions, and the image shows a standard football match scene without any clear emotional cues.	The image of the St Bede vs. St Augustines football match conveys a sense of camaraderie and sportsmanship. The players are engaged in a moment of play, suggesting a positive and supportive atmosphere. The neutral sentiment label aligns with the overall feeling of cambered athleticism and teamwork, without evoking strong emotions.	Neutral 😐 ✓
	MDCA	Because the writer is promoting the team's friendly football match with St. Albans, which suggests that they are a fan of the team. The use of the hashtag #sportmanship also suggests that the writer has a positive opinion of St. Bede. Overall, the language and tone of the tweet suggest that the sentiment towards St.Bede is positive.	It is a friendly football cup.	Positive 😀 ✗
	GPT-4o	Based on the image-text pair, the sentiment towards St Bede in the given text and image is Positive, because the text highlights a friendly and sportsmanlike attitude despite both teams being out of the competition, and the image shows players engaging in the match, fostering a positive environment.		Positive 😀 ✗
[Michael Oher]$_{NEG}$ is latest reminder that NFL pain lasts 24 / 7 / 365  Facial description: Michael Oher appears to have a serious or intense expression. The eyebrows are slightly furrowed, and the mouth is set in a straight line, which could indicate concentration, determination, or a readiness to face a challenge. There is no clear smile or frown.	Ground Truth	Based on the image-text pair, the sentiment towards Michael Oher in the given text and image is negative, because the text highlights the constant and enduring pain associated with his NFL career, implying a prolonged and difficult experience.	The image of Michael Oher in his NFL uniform, combined with the sentiment label "negative," conveys a sense of enduring struggle and fatigue. His posture and facial expression suggest a feeling of exhaustion or frustration, which aligns with the text's implication that the pain and challenges of an NFL career are relentless and unforgiving. The overall impression is one of continuous physical and emotional burden.	Negative ☹
	Chimera	The sentiment towards Michael Oher in the given text and image is negative, because the text highlights that his recent reminder that NFL pain lasts 24/7/365, which is a serious and distressing reminder.	The image of Michael Oher exudes a sense of determination and focus, aligning with the negative sentiment label. His serious expression and the context of NFL pain underscore a feeling of exhaustion and exhaustion, which complements the idea of prolonged pain and the accompanying reminder that "no matter how much you love it," evoking a negative feeling.	Negative ☹ ✓
	MDCA	Because the tweet does not express any clear positive or negative opinion towards him. The tweet simply states that he is a reminder that NFL pain lasts for a specific time period. There is no indication of any emotional attachment or opinion towards Michael Oher in the tweet, and the language used is informative rather than emotional. Therefore, the sentiment towards Michael Oer in this tweet is neutral.	He is reminder that NFL pain lasts 24 / 7 / 365.	Neutral 😐 ✗
	GPT-4o	Based on the image-text pair, the sentiment towards Michael Oher in the given text and image is Negative, because the text highlights the unending pain associated with NFL careers, using Michael Oher as an example, and alludes to the physical and possibly emotional toll he endures.		Negative ☹ ✓

Fig. 6. Three examples showcasing the predictions generated by Chimera, MDCA, and GPT-4o are presented for analysis. During the evaluation process, GPT-4o exclusively produces the semantic rationale (SR). The input image-text pair and auxiliary sentences are utilized solely by Chimera. For MDCA, the reasoning cause (RC), direct cause (DC), and sentiment prediction are derived through direct inference.

anced sentiment distributions. Specifically, attributes such as "visual" and "vibrant" positively contribute to sentiment analysis performance, whereas "focus" appears to significantly impair it. We speculate that since "focus" emphasizes specific image elements, potentially leads to an unbalanced interpretation of visual content. This localized emphasis can narrow the model's analytical scope, prioritizing details at the expense of broader context and compositional harmony. Consequently, the model may struggle to capture holistic aesthetic and emotional cues essential for accurate sentiment classification.

5.4 Comparison with Large Language Models

We evaluate the GPT-4o on the MASC task under a zero-shot setting. As shown in Table 7, GPT-4o achieves an accuracy of 46.87% and an F1 score of 47.47%, which is substantially lower than Chimera, which reports 81.61% accuracy and 77.98% F1 score. On the Twitter-2017 dataset, GPT-4o shows an improvement with an accuracy of 56.08% and an F1 score of 53.28%. However, this performance still trails behind Chimera, which reports 75.62% accuracy and 74.59% F1 score. Surprisingly, removing the image input results in an improvement in the model's accuracy and F1 score, reaching

TABLE 7
The experimental results (%) of GPT-4o on the MASC task under a zero-shot setting are presented. The best-performing results highlighted in **bold**. The term "dis" refers to the percentage of samples where the sentiment polarity associated with a specific aspect cannot be discerned.

Method	Twitter-2015			Twitter-2017		
	Acc	F1	Dis	Acc	F1	Dis
Chimera	**81.61**	**77.98**	-	**75.62**	**74.59**	-
GPT-4o	46.87	47.47	0.2	56.08	53.28	0.5
GPT-4o w/o image	67.02	62.38	-	59.64	60.35	-

67.02% and 62.38% on the Twitter-2015 dataset, respectively. This observation contrasts sharply with the phenomenon observed in the baseline model. Similarly, in the Twitter-2017 dataset, the performance of GPT-4o without image input is slightly better than with the image input. We speculate that in task-specific models, incorporating image data typically improves sentiment classification performance, as these models are fine-tuned to leverage multi-modal inputs effectively. However, in a zero-shot setting, GPT-4o operates based on its general pre-trained knowledge, which may not be fully optimized for combining textual and visual inputs for sentiment analysis. In this setting, adding image input may introduce noise rather than meaningful information. Moreover, GPT-4o has a low Dis value on both datasets, which slightly decreases to 0 when the image input is removed. This further suggests that the model's ability to distinguish sentiment polarity is, to a certain extent, influenced by the inclusion of the visual modality.

5.5 Case Study

An additional case study is performed to provide a more comprehensive evaluation of the effectiveness of the proposed Chimera model. Figure 6 illustrates three representative examples, each corresponding to positive, neutral, and negative samples, respectively. As illustrated in the first example, MDCA is the sole model to predict "Neutral" for the target "Joanne Stiger," whereas the other three models accurately predict "Positive". This result is primarily due to the RC and DC generated by MDCA, which lack the expression of positive or negative sentiment. Notably, the RC predominantly emphasizes the textual content, overlooking the joyful atmosphere conveyed through the image. In the second example, an intriguing observation is that the situation is the exact opposite of the previous case. Here, only Chimera correctly predicts the sentiment polarity of the specific target, "St. Bede" as "Neutral" whereas both GPT-4o and MDCA incorrectly classify it as "Positive". It is observed that the SR of GPT-4o and the RC of MDCA both convey a positive sentiment, largely due to an overinterpretation and extrapolation of the textual content. In contrast, Chimera demonstrates accurate prediction by appropriately integrating a balanced understanding of the image content and its aesthetic attributes. In the final example, both Chimera and GPT-4o accurately identify the sentiment polarity of "Michael Oher" as "Negative". MDCA's incorrect prediction of "Neutral" may be attributed to its generated RC and DC failing to account for the individual's expression, thereby overlooking critical semantic cues present in the visual content. With the aid of facial descriptions, Chimera effectively captures and aligns fine-grained emotional cues from visual content, enabling it to generate coherent SR and IR and achieve accurate predictions. The above representative instances further verify that incorporating cognitive and aesthetic sentiment causality enhances sentiment classification accuracy in MABSA.

6 Conclusion

In this paper, we propose a cognitive sentiment causality understanding framework tailored for multimodal aspect-based sentiment classification. The framework, which is novel in its approach, consists of four primary components: linguistic-aware semantic alignment, a translation module, rationale dataset construction, and rationale-aware learning. The linguistic-aware semantic alignment component facilitates visual patch-token level alignment through dynamic patch selection and semantic patch calibration. The translation module transforms holistic image and object-level visual information into corresponding emotion-laden textual representations. The rationale dataset construction involves designing refined prompts and leveraging LLMs to generate semantic and impression rationale. Finally, rationale-aware learning incorporates semantic explanations and affective-cognitive resonance to enhance the model's capacity to understand cognitive sentiment causality. Experimental results on three Twitter datasets demonstrate that the proposed Chimera achieves performance gains over SOTA baselines.

Acknowledgments

This research is supported by the Shanghai Science and Technology Innovation Action Plan (No. 24YF2710100), the Shanghai Special Project to Promote High-quality Industrial Development (No. RZ-CYAI-01-24-0288), the National Nature Science Foundation of China (No. 62477010), the Science and Technology Commission of Shanghai Municipality Grant (No. 22511105901, No. 21511100402), the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (STEM RIE2025 Award MOE-T2EP20123-0005) and by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore.

References

[1] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1743–1753, 2023.

[2] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1–42, 2024.

[3] R. Mao, M. Ge, S. Han, W. Li, K. He, L. Zhu, and E. Cambria, "A survey on pragmatic processing techniques," *Information Fusion*, vol. 114, p. 102712, 2025.

[4] R. Fan, S. Li, T. He, and Y. Liu, "Aspect-based sentiment analysis with syntax-opinion-sentiment reasoning chain," in *Proceedings of the 31st International Conference on Computational Linguistics*, 2025, pp. 3123–3137.

[5] L. Xiao, Y. Xue, H. Wang, X. Hu, D. Gu, and Y. Zhu, "Exploring fine-grained syntactic information for aspect-based sentiment classification with dual graph neural networks," *Neurocomputing*, vol. 471, pp. 48–59, 2022.

[6] Y. Ma, R. Mao, Q. Lin, P. Wu, and E. Cambria, "Quantitative stock portfolio optimization by multi-task learning risk and return," *Information Fusion*, vol. 104, p. 102165, 2024.

[7] K. Du, F. Xing, R. Mao, and E. Cambria, "FinSenticNet: A concept-level lexicon for financial sentiment analysis," in *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2023, pp. 109–114.

[8] X. Zhang, R. Mao, and E. Cambria, "SenticVec: Toward robust and human-centric neurosymbolic sentiment analysis," in *Findings of the Association for Computational Linguistics: ACL*. Association for Computational Linguistics, 2024, pp. 4851–4863.

[9] S. Zhao, M. Jia, L. A. Tuan, F. Pan, and J. Wen, "Universal vulnerabilities in large language models: Backdoor attacks for in-context learning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 11 507–11 522.

[10] J. Xu, L. Xiao, A. Wu, T. Ma, D. Dong, and L. He, "Bidirectional directed acyclic graph neural network for aspect-level sentiment classification," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 24, no. 4, pp. 1–14, 2025.

[11] J. Xu, S. Yang, L. Xiao, Z. Fu, X. Wu, T. Ma, and L. He, "Graph convolution over the semantic-syntactic hybrid graph enhanced by affective knowledge for aspect-level sentiment classification," in *2022 international joint conference on neural networks (IJCNN)*. IEEE, 2022, pp. 1–8.

[12] S. Zhao, J. Wen, A. Luu, J. Zhao, and J. Fu, "Prompt as triggers for backdoor attack: Examining the vulnerability in language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 12 303–12 317.

[13] J. YU and J. JIANG, "Adapting bert for target-oriented multimodal sentiment classification.(2019)," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 5408–5414.

[14] J. Yu, J. Jiang, and R. Xia, "Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 429–439, 2019.

[15] J. Yu, J. Wang, R. Xia, and J. Li, "Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching." in *IJCAI*, 2022, pp. 4482–4488.

[16] Y. Ling, J. Yu, and R. Xia, "Vision-language pre-training for multimodal aspect-based sentiment analysis," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 2149–2159.

[17] L. Yang, J.-C. Na, and J. Yu, "Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis," *Information Processing & Management*, vol. 59, no. 5, p. 103038, 2022.

[18] R. Zhou, W. Guo, X. Liu, S. Yu, Y. Zhang, and X. Yuan, "Aom: Detecting aspect-oriented information for multimodal aspect-based sentiment analysis," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 8184–8196.

[19] Z. Khan and Y. Fu, "Exploiting bert for multimodal target sentiment classification through input space translation," in *Proceedings of the 29th ACM international conference on multimedia*, 2021, pp. 3034–3042.

[20] L. Xiao, E. Zhou, X. Wu, S. Yang, T. Ma, and L. He, "Adaptive multi-feature extraction graph convolutional networks for multimodal target sentiment analysis," in *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2022, pp. 1–6.

[21] L. Xiao, X. Wu, S. Yang, J. Xu, J. Zhou, and L. He, "Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis," *Information Processing & Management*, vol. 60, no. 6, p. 103508, 2023.

[22] Y. Huang, Z. Chen, J. Chen, J. Z. Pan, Z. Yao, and W. Zhang, "Target-oriented sentiment classification with sequential cross-modal semantic graph," in *International Conference on Artificial Neural Networks*. Springer, 2023, pp. 587–599.

[23] Q. Wang, H. Xu, Z. Wen, B. Liang, M. Yang, B. Qin, and R. Xu, "Image-to-text conversion and aspect-oriented filtration for multimodal aspect-based sentiment analysis," *IEEE Transactions on Affective Computing*, 2023.

[24] L. Xiao, X. Wu, J. Xu, W. Li, C. Jin, and L. He, "Atlantis: Aesthetic-oriented multiple granularities fusion network for joint multimodal aspect-based sentiment analysis," *Information Fusion*, vol. 106, p. 102304, 2024.

[25] H. Yang, Y. Zhao, and B. Qin, "Face-sensitive image-to-emotional-text cross-modal translation for multimodal aspect-based sentiment analysis," in *Proceedings of the 2022 conference on empirical methods in natural language processing*, 2022, pp. 3324–3335.

[26] R. Fan, T. He, M. Chen, M. Zhang, X. Tu, and M. Dong, "Dual causes generation assisted model for multimodal aspect-based sentiment classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[27] J. Wang, Z. Li, J. Yu, L. Yang, and R. Xia, "Fine-grained multimodal named entity recognition and grounding with a generative framework," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3934–3943.

[28] X. Zhang, R. Mao, K. He, and E. Cambria, "Neurosymbolic sentiment analysis with dynamic word sense disambiguation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, 2023, pp. 8772–8783.

[29] Q. Lu, X. Sun, Y. Long, Z. Gao, J. Feng, and T. Sun, "Sentiment analysis: Comprehensive reviews, recent advances, and open challenges," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[30] H. Liu, W. Wang, and H. Li, "Interpretable multimodal misinformation detection with logic reasoning," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 9781–9796.

[31] R. Mao, K. Du, Y. Ma, L. Zhu, and E. Cambria, "Discovering the cognition behind language: Financial metaphor analysis with MetaPro," in *2023 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2023, pp. 1211–1216.

[32] E. Cambria, X. Zhang, R. Mao, M. Chen, and K. Kwok, "SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing," in *Proceedings of International Conference on Human-Computer Interaction (HCII)*, Washington DC, USA, 2024, pp. 197–216.

[33] K. Du, R. Mao, F. Xing, and E. Cambria, "Explainable stock price movement prediction using contrastive learning," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, Idaho, USA, 2024, pp. 529–537.

[34] H. Zhang, X. Zhou, Z. Shen, and Y. Li, "Privfr: Privacy-enhanced federated recommendation with shared hash embedding," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[35] E. Yang, L. Shen, G. Guo, X. Wang, X. Cao, J. Zhang, and D. Tao, "Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities," *arXiv preprint arXiv:2408.07666*, 2024.

[36] L. Xiao, R. Mao, X. Zhang, L. He, and E. Cambria, "Vanessa: Visual connotation and aesthetic attributes understanding network for multimodal aspect-based sentiment analysis," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 11 486–11 500.

[37] J. Kruk, J. Lubin, K. Sikka, X. Lin, D. Jurafsky, and A. Divakaran, "Integrating text and image: Determining multimodal document intent in instagram posts," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 4622–4632.

[38] H. Liu, W. Wang, and H. Li, "Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 4995–5006.

[39] R. Mao and X. Li, "Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 13 534–13 542.

[40] T. Yue, R. Mao, H. Wang, Z. Hu, and E. Cambria, "KnowleNet: Knowledge fusion network for multimodal sarcasm detection," *Information Fusion*, vol. 100, p. 101921, 2023.

[41] C. Fan, J. Lin, R. Mao, and E. Cambria, "Fusing pairwise modalities for emotion recognition in conversations," *Information Fusion*, vol. 106, p. 102306, 2024.

[42] L. Yang, Z. Wang, Z. Li, J.-C. Na, and J. Yu, "An empirical study of multimodal entity-based sentiment analysis with chatgpt: Improving in-context learning via entity-aware contrastive learning," *Information Processing & Management*, vol. 61, no. 4, p. 103724, 2024.

[43] L. Yang, J. Wang, J.-C. Na, and J. Yu, "Generating paraphrase sen-

tences for multimodal entity-category-sentiment triple extraction," *Knowledge-Based Systems*, vol. 278, p. 110823, 2023.

[44] J. Zhou, J. Zhao, J. X. Huang, Q. V. Hu, and L. He, "Masad: A large-scale dataset for multimodal aspect-based sentiment analysis," *Neurocomputing*, vol. 455, pp. 47–58, 2021.

[45] W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam, "A survey on aspect-based sentiment analysis: Tasks, methods, and challenges," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 11, pp. 11 019–11 038, 2022.

[46] X. Ju, D. Zhang, R. Xiao, J. Li, S. Li, M. Zhang, and G. Zhou, "Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection," in *Proceedings of the 2021 conference on empirical methods in natural language processing*, 2021, pp. 4395–4405.

[47] J. Mu, F. Nie, W. Wang, J. Xu, J. Zhang, and H. Liu, "Mocolnet: A momentum contrastive learning network for multimodal aspect-level sentiment analysis," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[48] F. Zhao, C. Li, Z. Wu, Y. Ouyang, J. Zhang, and X. Dai, "M2df: Multi-grained multi-curriculum denoising framework for multi-modal aspect-based sentiment analysis," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 9057–9070.

[49] E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho, "Seven pillars for the future of artificial intelligence," *IEEE Intelligent Systems*, vol. 38, no. 6, pp. 62–69, 2023.

[50] R. Arnheim, *Art and visual perception: A psychology of the creative eye*. Univ of California Press, 1954.

[51] V. S. Ramachandran and W. Hirstein, "The science of art: A neurological theory of aesthetic experience," *Journal of consciousness Studies*, vol. 6, no. 6-7, pp. 15–51, 1999.

[52] H. Zeng, Z. Cao, L. Zhang, and A. C. Bovik, "A unified probabilistic formulation of image aesthetic assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 1548–1561, 2019.

[53] G. C. Cupchik and J. László, *Emerging visions of the aesthetic process: In psychology, semiology, and philosophy*. Cambridge University Press, 1992.

[54] X. Jin, L. Wu, G. Zhao, X. Li, X. Zhang, S. Ge, D. Zou, B. Zhou, and X. Zhou, "Aesthetic attributes assessment of images," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 311–319.

[55] J. Ke, K. Ye, J. Yu, Y. Wu, P. Milanfar, and F. Yang, "Vila: Learning image aesthetics from user comments with vision-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 041–10 051.

[56] J. Kruk, C. Ziems, and D. Yang, "Impressions: Visual semiotics and aesthetic impact understanding," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 12 273–12 291.

[57] R. Anil, A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, E. Taropa, P. Bailey, Z. Chen *et al.*, "Palm 2 technical report," *arXiv preprint arXiv:2305.10403*, 2023.

[58] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, "GPTEval: A survey on assessments of ChatGPT and GPT-4," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, 2024, pp. 7844–7866.

[59] S. Zhao, L. A. Tuan, J. Fu, J. Wen, and W. Luo, "Exploring clean label backdoor attacks and defense in language models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[60] S. Zhao, X. Xu, L. Xiao, J. Wen, and L. A. Tuan, "Clean-label backdoor attack and defense: An examination of language model vulnerability," *Expert Systems with Applications*, vol. 265, p. 125856, 2025.

[61] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[62] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, K. Millican *et al.*, "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[63] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[64] H. Liu, W. Wang, H. Sun, A. Rocha, and H. Li, "Robust domain misinformation detection via multi-modal feature alignment," *IEEE Transactions on Information Forensics and Security*, 2023.

[65] R. Mao, K. He, C. Ong, Q. Liu, and E. Cambria, "Metapro 2.0: Computational metaphor processing on the effectiveness of anomalous language modeling," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 9891–9908.

[66] Z. Tan, D. Li, S. Wang, A. Beigi, B. Jiang, A. Bhattacharjee, M. Karami, J. Li, L. Cheng, and H. Liu, "Large language models for data annotation: A survey," *arXiv preprint arXiv:2402.13446*, 2024.

[67] R. Mao, G. Chen, X. Li, M. Ge, and E. Cambria, "A comparative analysis of metaphorical cognition in chatgpt and human minds," *Cognitive Computation*, vol. 17, no. 1, p. 35, 2025.

[68] Y. Jia, X. Wu, H. Li, Q. Zhang, Y. Hu, S. Zhao, and W. Fan, "Uni-retrieval: A multi-style retrieval framework for stem's education," *arXiv preprint arXiv:2502.05863*, 2025.

[69] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[70] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[71] P. Wang, A. Chan, F. Ilievski, M. Chen, and X. Ren, "Pinto: Faithful language reasoning using prompt-generated rationales," in *The Eleventh International Conference on Learning Representations*, 2023.

[72] P. Wang, Z. Wang, Z. Li, Y. Gao, B. Yin, and X. Ren, "Scott: Self-consistent chain-of-thought distillation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 5546–5558.

[73] H. Liu, Z. Teng, L. Cui, C. Zhang, Q. Zhou, and Y. Zhang, "Logicot: Logical chain-of-thought instruction tuning," in *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[74] M. Kang, S. Lee, J. Baek, K. Kawaguchi, and S. J. Hwang, "Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[75] Y. Li, A. Dao, W. Bao, Z. Tan, T. Chen, H. Liu, and Y. Kong, "Facial affective behavior analysis with instruction tuning," in *European Conference on Computer Vision*. Springer, 2025, pp. 165–186.

[76] J. Guo, J. Deng, A. Lattas, and S. Zafeiriou, "Sample and computation redistribution for efficient face detection," in *International Conference on Learning Representations*, 2021.

[77] S. Wiegreffe, J. Hessel, S. Swayamdipta, M. Riedl, and Y. Choi, "Reframing human-ai collaboration for generating free-text explanations," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 632–658.

[78] L. Meng, H. Li, B.-C. Chen, S. Lan, Z. Wu, Y.-G. Jiang, and S.-N. Lim, "Adavit: Adaptive vision transformers for efficient image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 309–12 318.

[79] Z. Fu, L. Zhang, H. Xia, and Z. Mao, "Linguistic-aware patch slimming framework for fine-grained cross-modal alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 26 307–26 316.

[80] C. Maddison, A. Mnih, and Y. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *Proceedings of the international conference on learning Representations*. International Conference on Learning Representations, 2017.

[81] Z. Zong, K. Li, G. Song, Y. Wang, Y. Qiao, B. Leng, and Y. Liu, "Self-slimmed vision transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 432–448.

[82] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," *arXiv preprint arXiv:1707.05612*, 2017.

[83] L. Yang, J. Yu, C. Zhang, and J.-C. Na, "Fine-grained sentiment analysis of political tweets with entity-aware multimodal network," in *Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part I 16*. Springer, 2021, pp. 411–420.

[84] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[85] I. Loshchilov, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
[86] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[87] D. Tang, B. Qin, and T. Liu, "Aspect level sentiment classification with deep memory network," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 214–224.
[88] F. Fan, Y. Feng, and D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 3433–3442.
[89] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1. Minneapolis, Minnesota, 2019, p. 2.
[90] N. Xu, W. Mao, and G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 371–378.
[91] J. Yu, K. Chen, and R. Xia, "Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1966–1978, 2022.
[92] D. Liu, L. Li, X. Tao, J. Cui, and Q. Xie, "Descriptive prompt paraphrasing for target-oriented multimodal sentiment classification," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 4174–4186.
[93] B. Yang and J. Li, "Visual elements mining as prompts for instruction learning for target-oriented multimodal sentiment classification," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 6062–6075.
[94] J. Camacho-Collados, K. Rezaee, T. Riahi, A. Ushio, D. Loureiro, D. Antypas, J. Boisson, L. E. Anke, F. Liu, and E. Martínez-Cámara, "Tweetnlp: Cutting-edge natural language processing for social media," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2022, pp. 38–49.
[95] J. Ye, J. Zhou, J. Tian, R. Wang, Q. Zhang, T. Gui, and X.-J. Huang, "Rethinkingtmsc: An empirical study for target-oriented multimodal sentiment classification," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 270–277.
[96] M. Ivanova and S. French, *The aesthetics of science: beauty, imagination and understanding*. Routledge, 2020.

Luwei Xiao is currently pursuing his Ph.D. degree in the School of Computer Science and Technology at East China Normal University, Shanghai, China, under the supervision of Prof. Liang He. He is presently conducting an academic visit to the College of Computing and Data Science at Nanyang Technological University, Singapore, under the supervision of Prof. Erik Cambria, with funding support from the China Scholarship Council (CSC). His research interests encompass multimodal learning, sentiment analysis, and image aesthetic assessment.

Rui Mao is a Research Scientist and Lead Investigator at Nanyang Technological University. He obtained his Ph.D. degree in Computing Science from the University of Aberdeen. His research interest lies in NLP, cognitive computing, and their applications in finance and cognitive science. He and his funded company (Ruimao Tech) have developed an end-to-end system (MetaPro) for computational metaphor processing and a neural search engine (wensousou.com) for searching Chinese ancient poems with modern language. He served as Area Chair in COLING and EMNLP and Associate Editor in IEEE Transactions on Affective Computing, Expert Systems, Information Fusion and Neurocomputing. Contact him at rui.mao@ntu.edu.sg.

Shuai Zhao obtained his Ph.D. degree from Jinan University in 2024. He spent one year as a visiting student and six months as a research assistant at the School of Computer Science and Engineering, Nanyang Technological University. He is now a Postdoctoral Researcher at the College of Computing and Data Science, Nanyang Technological University. His current research interests include deep learning and natural language processing for code generation, summary generation, text classification and backdoor attacks.

Qika Lin received his Ph.D. degree at Xi'an Jiaotong University. Currently, he is a Research Fellow at the National University of Singapore. His research interests include natural language processing, knowledge reasoning, and multimodal learning. He has published papers in top-tier journals/conferences, including TKDE, ACL, SIGIR, KDD, ICDE, and IJCAI. He has actively contributed to several journals/conferences as a reviewer or PC member, including TPAMI, IJCV, TKDE, TMC, TNNLS, NeurIPS, ICLR, SIGIR, ACL, and EMNLP. He also served as a Guest Editor of IEEE TCSS and Information Fusion.

Yanhao Jia is a phd student at Nanyang Technological University. He obtained his bechealor degree in Computing Science from Shandong University. He has published over seven conference/journal papers on ECCV/NeurIPS/IEEE Trans on nuclear science and been the reviewer for ACM MM and ECCV.

Liang He received his PhD degree from the Department of Computer Science and Technology, East China Normal University, China. He is now a professor and the Vice Dean of the School of Computer Science and Technology, East China Normal University. His current research interest includes Natural Language Processing, Knowledge Processing, and Human in the Loop for Decision-making.

Erik Cambria is a Professor at Nanyang Technological University, where he also holds the appointment of Provost Chair in Computer Science and Engineering, and a Visiting Professor at MIT Media Lab. He is also a founder of several AI companies, such as SenticNet, offering B2B sentiment analysis services, and finaXai, providing fully explainable financial insights. His research focuses on neurosymbolic AI for interpretable, trustworthy, and explainable affective computing in domains like social media monitoring, financial forecasting, and AI for social good. He is an IEEE Fellow, Associate Editor of various top-tier AI journals, e.g., Information Fusion and IEEE Transactions on Affective Computing, and is involved in several international conferences as keynote speaker, program chair and committee member. Contact him at cambria@ntu.edu.sg.