Beyond Words: Enhancing Desire, Emotion, and Sentiment Recognition with Non-Verbal Cues

Wei Chen
College of Informatics, Huazhong Agricultural University
Wuhan, China
weichen5498@webmail.hzau.edu.cn

Tongguan Wang
College of Informatics, Huazhong Agricultural University
Wuhan, China
wang_tg@webmail.hzau.edu.cn

Feiyue Xue
College of Informatics, Huazhong Agricultural University
Wuhan, China
xuefeiyue@webmail.hzau.edu.cn

Junkai Li
College of Informatics, Huazhong Agricultural University
Wuhan, China
junkaili@webmail.hzau.edu.cn

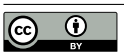
Hui Liu
College of Informatics, Huazhong Agricultural University
Wuhan, China
liuhui_1003@webmail.hzau.edu.cn

Ying Sha*†
College of Informatics, Huazhong Agricultural University
Engineering Research Center of Intelligent Technology for Agriculture
Wuhan, China
shaying@mail.hzau.edu.cn

Abstract

Multimodal desire understanding, a task closely related to both emotion and sentiment that aims to infer human intentions from visual and textual cues, is an emerging yet underexplored task in affective computing with applications in social media analysis. Existing methods for related tasks predominantly focus on mining verbal cues, often overlooking the effective utilization of non-verbal cues embedded in images. To bridge this gap, we propose a ***Sy**mmetrical Bidirectional Multimodal Learning Framework for **D**esire, **E**motion, and **S**entiment Recognition* (SyDES). The core of SyDES is to achieve bidirectional fine-grained modal alignment between text and image modalities. Specifically, we introduce a mixed-scaled image strategy that combines global context from low-resolution images with fine-grained local features via masked image modeling (MIM) on high-resolution sub-images, effectively capturing intention-related visual representations. Then, we devise symmetrical cross-modal decoders, including a text-guided image decoder and an image-guided text decoder, which enable mutual reconstruction and refinement between modalities, facilitating deep cross-modal interaction. Furthermore, a set of dedicated loss functions is designed to harmonize potential conflicts between the MIM and modal alignment objectives during optimization. Extensive evaluations on the MSED benchmark demonstrate the superiority of our approach, which establishes a new state-of-the-art performance with 1.1% F1-score improvement in desire understanding. Consistent gains in emotion and sentiment recognition further validate its generalization ability and the necessity of utilizing non-verbal cues. Our code is available at: https://github.com/especiallyW/SyDES.

*Corresponding author.
†Also with Hubei Engineering Technology Research Center of Agricultural Big Data, Key Laboratory of Smart Farming for Agricultural Animals.



WWW '26, Dubai, United Arab Emirates.

ACM ISBN 979-8-4007-2307-0/2026/04
https://doi.org/10.1145/3774904.3792393

CCS Concepts

• **Computing methodologies** → **Natural language processing**.

Keywords

Social Media; Human Desire Understanding; Sentiment Analysis; Emotion Recognition; Multimodal Learning; Multimodal Fusion

ACM Reference Format:
Wei Chen, Tongguan Wang, Feiyue Xue, Junkai Li, Hui Liu, and Ying Sha. 2026. Beyond Words: Enhancing Desire, Emotion, and Sentiment Recognition with Non-Verbal Cues. In *Proceedings of the ACM Web Conference 2026 (WWW '26), April 13–17, 2026, Dubai, United Arab Emirates.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3774904.3792393

1 Introduction

Human Desire is a fundamental intention that reflects a strong wish for certain objects or states [25]. It interacts closely with emotion and sentiment, shaping affective life experiences while being modulated by them in return. Such three tasks form interconnected and essential components of the human experience, driving our actions and decisions. For example, in Figure 1 (a), a couple smiling and preparing food in a kitchen can be interpreted as expressing a "***romance***" desire, which explains positive sentiment and happy emotion. Figure 1 (b) depicts a man's exaggerated movements to avoid security cameras. It can be understood as fear and negative sentiment driven by a desire for safety and privacy. Therefore, if a machine were capable of accurately inferring such desire intents, it would move research closer to recognizing human emotional intelligence [9]. However, the specific problem of desire understanding has received comparatively little dedicated attention.

A key to desire understanding lies in effectively leveraging the non-verbal cues embedded in images, which provide rich contextual information beyond textual descriptions. However, existing methods in multimodal emotion and sentiment recognition [39, 45] still predominantly focus on enhancing verbal cues (e.g., via graph neural networks), treating image-based non-verbal cues merely as auxiliary features to be extracted and fused superficially. This underutilization constitutes a fundamental limitation. In practice,

non-verbal cues play a crucial role. For instance, the grimacing expression in Figure 1 (c) could indicate disgust toward broccoli or be part of a playful interaction with family; without the rich context from the image, text-based inference is inherently ambiguous. We argue that overcoming this limitation is not only beneficial for emotion and sentiment analysis but is particularly critical for advancing the more complex task of desire understanding.

To this end, we propose SyDES, *a **Sy**mmetrical Bidirectional Multimodal Learning Framework for **D**esire, **E**motion, and **S**entiment recognition*. The framework emphasizes deep utilization of non-verbal visual cues while ensuring that verbal cues remain effectively exploited. Specifically, the input image is processed as both a low-resolution version and a high-resolution version using a shared image encoder. The low-resolution image provides global visual representations for cross-modal alignment. The high-resolution image is processed by mixed-scale image strategy to get high-resolution sub-images, and these sub-images are modeled with masked image modeling to encourage the encoder to better learn fine-grained local features. A text-guided image decoder is introduced so that image reconstruction can be guided by textual semantics. Conversely, an image-guided text decoder is employed so that text decoding can incorporate multi-scale visual information. In the meantime, we design a set of loss functions (e.g., local-global semantic similarity loss, cross-modal feature-distribution consistency loss) to allow reconstructed image to maintain modal alignment of fine-grained local visual features and global visual representations, thereby avoid over-reliance on specific regional features. These mechanisms enable mutual guidance and semantic alignment of textual representations, local visual features and global visual features. The fused text outputs are then passed to a lightweight multi-layer perception (MLP) for downstream prediction (see Figure 2 and Section 3 for details).

To evaluate our approach, extensive experiments are conducted on MSED dataset, the first multimodal benchmark encompassing desire understanding, emotion recognition, and sentiment analysis. Our approach is primarily evaluated on the desire understanding task, where it achieves a significant improvement of 1.1% in F1-score, establishing a new state-of-the-art. Furthermore, consistent performance gains of 0.6% in emotion recognition and 0.9% in sentiment analysis demonstrate the framework's generalizability and underscore the necessity of fully utilizing non-verbal cues. Our contributions can be summarized:

(1) We propose SyDES, a novel framework by deeply leveraging non-verbal visual cues through a symmetrical bidirectional architecture.
(2) We introduce a mixed-scale image strategy and symmetrical cross-modal decoders, to capture fine-grained features and deep cross-modal interaction between text and image modalities.
(3) We design a set of dedicated loss functions to harmonize the objectives of MIM and modal alignment, ensuring consistent learning across different modalities and scales.
(4) We provide comprehensive experiments and ablation studies on the MSED dataset, validating the effectiveness and generalization of our proposed SyDES for multimodal desire understanding, emotion recognition, and sentiment analysis.

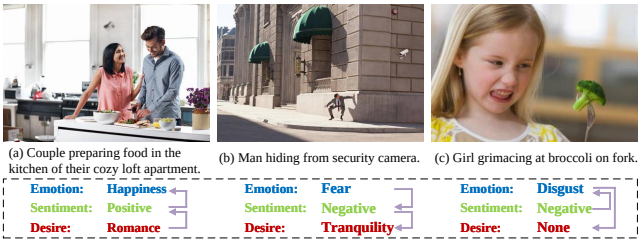


Figure 1: Examples of multimodal desire, emotion, and sentiment.

2 Related Work

2.1 Sentiment Analysis and Emotion Recognition

2.1.1 Text-only sentiment analysis. Sentiment analysis [15, 23, 30, 32] has long been a central topic in natural language processing. Early studies were largely unimodal. For example, Taboada et al. [31] proposed the lexicon-based Semantic Orientation CALculator, which leverages polarity-annotated lexica with negation handling. Pang et al. [23] and Gamallo et al. [5] applied classical classifiers (SVM, Naïve Bayes) to sentiment tasks. Kim et al. [15] introduced convolutional neural networks for text classification, while Tai et al. [32] proposed Tree-LSTM to model syntactic hierarchies. Yang et al. [40] developed the Hierarchical Attention Network (HAN) to select salient words and sentences for document-level classification. Wang et al. [34] incorporated attention into LSTM for aspect–sentiment modeling. More recently, Singh et al. [30] employed BERT for sentiment analysis on COVID-related tweets. These methods focus on textual signals; however, social media content often pairs text with images, and multimodal cues typically provide complementary information that improves predictive accuracy.

2.1.2 Multimodal sentiment analysis and emotion recognition. Multimodal sentiment classification [12, 37, 42] has attracted increasing interest for jointly modeling sentiment across modalities. You et al. [42] introduced cross-modal consistency regression (CCR) to fuse image and text features. Xu et al. [36] leveraged scene-level visual cues with attention to identify salient textual elements. Hu et al. [11] investigated users' latent affective states. From the interaction perspective, Xu et al. [37] explored iterative image–text relations; Huang et al. [12] proposed hybrid fusion of unimodal and cross-modal representations. Li et al. [17] used contrastive learning and augmentation to align token-level image–text features. Multimodal emotion recognition [16, 27, 43], which targets finer affective states, has also been extensively studied. Guo et al. [6] introduced multimodal news datasets and a layout-driven network; Nemaiti et al. [21] proposed a hybrid latent-space fusion method; Zhang et al. [43] combined manifold learning with deep convolutional networks; Xu et al. [35] used image captions as semantic cues; Yang et al. [38] employed memory and multi-view attention for integration. Despite these advances, most of the existing approaches simply leveraged holistic or local features extracted from different modalities to predict multimodal sentiments, which leads to suboptimal performance.

The emergence of graph neural networks (GNNs) [19, 28] enabled structured relation mining among verbal cues. Yang et al. [39] introduced a multi-channel GNN to capture global emotional attributes. Zhang et al. [44] proposed a multi-task interactive graph-attention network with local–global context modules. Wang et al. [33] enriched text representations with contextual world knowledge from large multimodal models. However, these approaches predominantly focus on mining verbal cues and often underutilize the rich information contained in non-verbal cues.

Motived by the importance of non-verbal cues, we hope to attain fine-gained features from image contextual information. Masked image modeling paradigm introduced by He et al. [7] has shown that an encoder can be encouraged, via a reconstruction-based self-supervised objective, to learn richer local representations. This insight motivates our proposal of a symmetrical bidirectional multimodal learning framework to ensure more effective exploitation of image-based non-verbal cues for desire understanding, emotion recognition, and sentiment analysis.

2.2 Multimodal Desire Understanding

Multimodal desire understanding concerns recognizing desires or intentions expressed in both textual and visual expression, and it remains an underexplored problem. Existing automated analyses of desire largely originate from psychology and philosophy. Lim et al. [18] developed a desire-understanding system based on four emotional states in audio and gestural cues. Cacioppo et al. [2] designed a multi-level kernel-density fMRI analysis to investigate differences and correlations between sexual desire and love. Schutte et al. [29] conducted a meta-analysis on 2,692 participants to examine links between curiosity and creativity. Hoppe et al. [10] estimated different levels of curiosity using eye-movement data and SVM. Yavuz et al. [41] proposed a data-mining approach for desire and intent using neural networks and Bayesian networks. Chauhan et al. [3] presented a multi-task multimodal deep attentive framework for offense, motivation, and sentiment analysis. Nevertheless, these researches commonly lack support for large-scale multimodal social media data and often do not fully exploit both visual and textual channels.

The recent introduction of the MSED dataset [14], the first multimodal dataset for desire understanding, provides a valuable benchmark. And subsequent work like MMTF-DES [1] attempts to improve performance through model ensemble. Nevertheless, such ensemble strategies incur significant computational costs and still lack a principled approach for deep, fine-grained cross-modal interaction. To address these limitations, we propose a novel symmetrical bidirectional framework that systematically leverages non-verbal cues through a mixed-scale image strategy and symmetrical decoders, offering an efficient and effective solution for multimodal desire understanding.

3 SyDES

Figure 2 illustrates the overall architecture of SyDES. Motivated by the necessity to deeply leverage non-verbal visual cues for desire understanding, our framework is built upon a mixed-scale image strategy to capture both global context and fine-grained local features. We first detail this strategy, followed by descriptions of the image encoder, text encoder, symmetrical decoders, the loss functions designed to harmonize different learning objectives, and the two-stage training paradigm.

3.1 Mixed-Scale Image Strategy

A core challenge in multimodal desire understanding is to extract comprehensive intention-related representations from images, which necessitates the model's ability to perceive both global context and fine-grained local details. High-resolution images can capture richer region-level features and local details, but processing them entirely at full resolution (e.g., 448×448) is computationally prohibitive. Conversely, conventional low-resolution inputs (e.g., 224×224) sacrifice crucial local information.

To balance perceptual granularity and computational efficiency, we adopt a mixed-scale image strategy. Given a high-resolution image I_i (e.g., 448×448) from batch size N, we generate one downsampled low-resolution image and partition the original image into four non-overlapping high-resolution sub-images. This yields five 224×224 images per original image, providing one global view and four detailed local views. Formally:

$$\begin{aligned} I_i^{(\mathrm{g})} &= \mathrm{DownSample}(I_i), \quad & I_i^{(\mathrm{g})} &\in \mathbb{R}^{224\times224\times3} \\ I_{i,n}^{(\mathrm{l})} &= \mathrm{Crop}_n(I_i), & I_{i,n}^{(\mathrm{l})} &\in \mathbb{R}^{224\times224\times3},\ n = 1,2,3,4 \end{aligned} \tag{1}$$

where DownSample uses bilinear interpolation, and Crop_n denotes corner or predefined cropping. Each 224×224 image is treated as an independent input to the image encoder to obtain its representations. This strategy efficiently provides the model with both global and local visual cues essential for understanding nuanced desires.

3.2 Model Architecture

Our proposed SyDES is built upon a symmetrical bidirectional architecture designed to achieve deep, bidirectional fusion between verbal and non-verbal cues. As illustrated in Figure 2, the framework comprises four core modules: image encoder, text encoder, text-guided image decoder, and image-guided text decoder, plus a lightweight MLP for downstream prediction.

3.2.1 Image Encoder. We employ a vision transformer [4] as the shared image encoder to process multi-scale visual representations. For the global context, we use the low-resolution image $I_i^{(\mathrm{g})}$, while for local details, we process high-resolution sub-images $I_{i,n}^{(\mathrm{l})}$ through masked image modeling. Specifically, each 224×224 image is divided into P patches and embedded as:

$$\begin{aligned} P_{\mathrm{emb},i,n}^{(\mathrm{g})} &= \left[v_i^{\mathrm{cls}}, v_i^1, \ldots, v_i^P\right] \in \mathbb{R}^{(P+1)\times C_1} \\ P_{\mathrm{emb},i,n}^{(\mathrm{l})} &= \left[v_{i,n}^{\mathrm{cls}}, v_{i,n}^1, \ldots, v_{i,n}^P\right] \in \mathbb{R}^{(P+1)\times C_1} \end{aligned}$$

where C_1 denotes the image embedding dimension and v^{cls} is the CLS token. For high-resolution sub-images $I_{i,n}^{(\mathrm{l})}$, we apply a binary mask vector $m_{i,n} \in \{0,1\}^P$ with $m_{i,n}[P] = 1$ indicating that patch p is masked. Following [7], we set the mask ratio to $m \in [0,1]$ and set the keep ratio to $r = 1 - m$. The set of unmasked indices is $M_{i,n} = \{p \mid m_{i,n}[p] = 0\}$ with $|M_{i,n}| = rP$. Selecting the unmasked tokens yields:

$$\widetilde{P}_{\mathrm{emb},i,n}^{(\mathrm{l})} = \mathrm{Select}\left(P_{\mathrm{emb},i,n}^{(\mathrm{l})}, M_{i,n}\right) \in \mathbb{R}^{(rP+1)\times C_1}$$

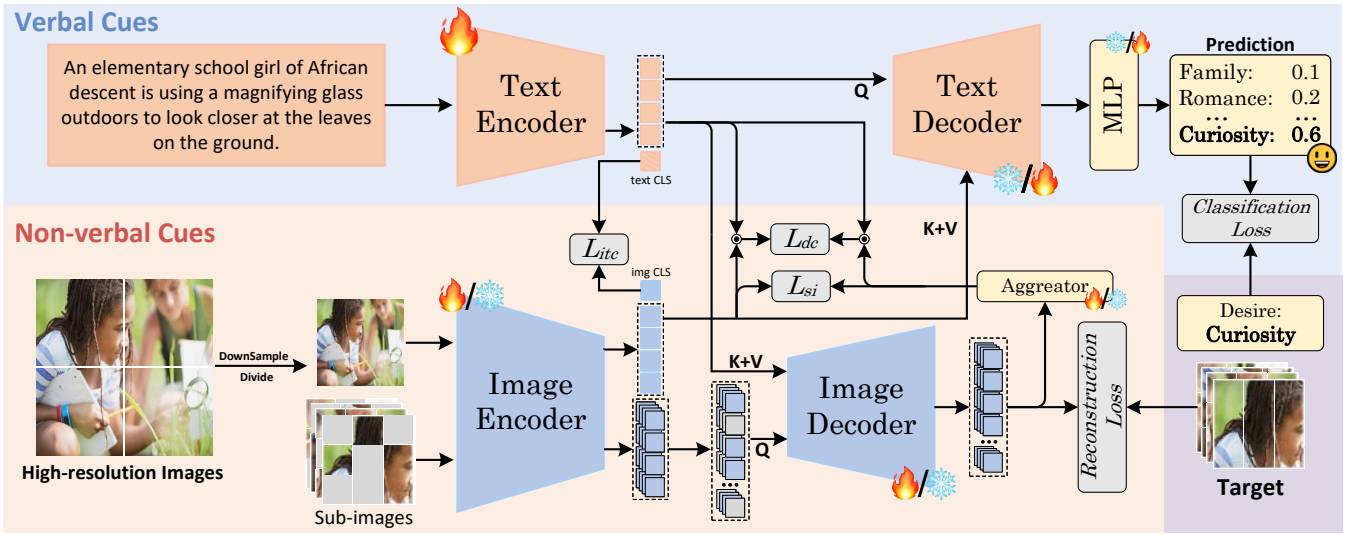


Figure 2: Overall architecture of SyDES. The model consists of four core modules: an image encoder, a text encoder, a text-guided image decoder, and an image-guided text decoder.

Both $P^{(g)}_{emb,i}$ and $\widetilde{P}^{(l)}_{emb,i,n}$ are fed into the image encoder to produce $V^{(g)}_i \in \mathbb{R}^{(P+1)\times C_1}$ and $V^{(l)}_{i,n} \in \mathbb{R}^{(rP+1)\times C_1}$, where the first row of $V^{(g)}_i$ corresponds to the low-resolution CLS visual feature v^{cls}_i.

3.2.2 Text Encoder. We adopt a casual masked transformer to model text inputs. Texts are tokenized into an embedding sequence of length $S+1$ represented as $W_{emb,i}$, and attain encoded features:

$$W_i = [w^1_i, \ldots, w^S_i, w^{cls}_i] \in \mathbb{R}^{(S+1)\times C_2}$$

with w^{cls} being the CLS text feature and C_2 denotes the text embedding dimension.

3.2.3 Text-guided Image Decoder. To recover masked patches from $\widetilde{P}^{(l,n)}_{emb,i}$ and make reconstruction aware of textual semantics, we adopt a text-guided imaged decoder. This process ensures that the image reconstruction is *context-aware*, refining the visual features based on verbal descriptions. Specifically, we project $V^{(l)}_{i,n}$ to the text embedding space via $\Pi_{v\to t} : \mathbb{R}^{C_1} \to \mathbb{R}^{C_2}$, and introduce mP learnable mask tokens $P_{mask} \in \mathbb{R}^{mP\times C_2}$. The decoder input for local crop n is:

$$D^{in}_{i,n} = [P_{mask};\ \Pi_{v\to t}(V^{(l)}_{i,n})] \in \mathbb{R}^{(P+1)\times C_2}$$

To stabilize fuse textual context, a gate-based fusion mechanism is used:

$$U^{img}_{i,n} = \mathrm{Gate}_{img}(W_i, D^{in}_{i,n}) \in \mathbb{R}^{(P+1)\times C_2}$$

As utilizing $D^{in}_{i,n}$ as the query and $U^{img}_{i,n}$ as the key and value, we leverage textual semantic information to compel the masked image reconstruction process to perceive verbal cues, and attained decode:

$$\{\widetilde{X}^p_{i,n}\}_{p\in\overline{M}_{i,n}},\ z_{i,n} = \mathrm{ImgDec}(D^{in}_{i,n},\ U^{img}_{i,n}),$$

where $\widetilde{X}^p_{i,n} \in \mathbb{R}^{(P+1)\times C_2}$ is the predicted pixel value for the p-th masked patch, and $z_{i,n} \in \mathbb{R}^{C_2}$ is a sub-image-level representation used for fine-grained alignment and revision. Through this *text-guided reconstruction*, the image features are refined to be semantically consistent with the accompanying text.

3.2.4 Image-guided Text Decoder. Symmetrically, we design an image-guided text decoder aims to refine the text representation by incorporating multi-scale non-verbal visual cues. It allows the text semantics to be grounded in and disambiguated by the rich context provided by the image, addressing the inherent ambiguity of text-only analysis.

We concatenate multi-scale visual features from the global image and all local sub-images:

$$V^{all}_i = [\Pi_{v\to t}(V^{(g)}_i),\ \Pi_{v\to t}(V^{(l)}_{i,1}), \ldots, \Pi_{v\to t}(V^{(l)}_{i,n})] \in \mathbb{R}^{nP\times C_2}$$

as the key and value, and use non-CLS text tokens $W^{1:S}_i$ as the query. A cross-attention mechanism then fuses these non-verbal visual cues into the text representation:

$$\widetilde{W}_i = \mathrm{TextDec}(W^{1:S}_i,\ V^{all}_i)$$

The output $\widetilde{W}_i \in \mathbb{R}^{S\times C_2}$ is a visually-refined text feature, which is then passed through a lightweight MLP for the final prediction $\widehat{y}_i$. This step constitutes the *image-guided refinement* of the text modality.

3.3 Loss Functions

3.3.1 Reconstruction Loss. The reconstruction loss supervises the text-guided image decoder to recover masked patches in high-resolution sub-images. By minimizing the *mean squared error* between the the masked image tokens and the reconstructed tokens, this loss enhances the model's capability to capture fine-grained visual details:

$$\mathcal{L}_{rec} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{\sum_n |\overline{M}_{i,n}|}\sum_n\sum_{p\in\overline{M}_{i,n}} \|X^p_{i,n} - \widetilde{X}^p_{i,n}\|_2\right) \quad (2)$$

where $X_{i,n}^p$ and $\widetilde{X}_{i,n}^p$ are the RGB pixel value and predicted pixel values for the p-th masked patch in the n-th sub-image, $|\overline{M}_{i,n}|$ denotes the number of masked patches.

3.3.2 Image-Text Contrastive Loss. To align the global representations of image and text modalities in a shared semantic space, we employ an image-text contrastive loss. This loss encourages matched image-text pairs to have similar representations while pushing unmatched pairs apart:

$$\begin{aligned}\mathcal{L}_{itc} = &\frac{1}{2N}\left[\sum_{i=1}^{N} -\log\left(\frac{\exp\left(\langle v, w\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle v, w_j\rangle/\tau\right)}\right)\right] \\ &+\frac{1}{2N}\left[\sum_{i=1}^{N} -\log\left(\frac{\exp\left(\langle w, v\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle w, v_j\rangle/\tau\right)}\right)\right]\end{aligned} \tag{3}$$

where $v = v_i^{cls}$ and $w = w_i^{\text{cls}}$ are the normalized global features from the low-resolution image and text, respectively, $\langle\cdot,\cdot\rangle$ denotes the inner product, and τ is a temperature parameter.

3.3.3 Local-Global Semantic Similarity Loss. It may cause the reconstructed features to deviate from global visual semantics, although MIM focuses on local details. We introduce a cross-scale consistency loss to maintain alignment between local reconstructed features and global visual representations, resulting in over rely local information. To enforce consistency between the reconstructed local features and the global visual representations, we perform learnable weighted aggregation on the sub-image-level global representation $\{z_{i,n}\}_{n=1}^4$ to obtain:

$$\begin{aligned} e_{i,n} &= u^T \tanh\left(z_{i,n}W_z + b\right) \\ \alpha_{i,n} &= \left[\text{softmax}\left(\left[e_{i,1},\ldots,e_{i,4}\right]\right)\right]_n \\ P_{\text{agg},i} &= \text{MLP}\left(\sum_{n=1}^{4}\alpha_{i,n}z_{i,n}\right)\end{aligned}$$

here, W_z, u, b are learnable parameters, and the MLP projects the aggregated feature into same semantic space as v_i^{cls}. After normalizing to all features, we compute:

$$\mathcal{L}_{si} = \frac{1}{N}\sum_{i=1}^{N}\|v_i^{\text{cls}} - P_{\text{agg},i}\|_2^2 \tag{4}$$

This loss ensures that the reconstructed local pixel features from high-resolution sub-images remain consistent with the global image semantics from low-resolution images, achieving a balance between local and global features.

3.3.4 Cross-Modal Feature-Distribution Consistency Loss. The pixel-level reconstruction ($\mathcal{L}_{rec}$) and semantic-level alignment ($\mathcal{L}_{itc}$) objectives may conflict. To harmonize them, we enforce distribution consistency between the text-to-reconstructed-image and text-to-global-image similarity distributions:

$$\mathcal{L}_{dc} = \frac{1}{N}\sum_{i=1}^{N}\left[\text{KL}(S_{t2l}\|S_{t2g}) + \lambda\cdot\text{H}(S_{t2l})\right] \tag{5}$$

where $S_{t2l} = S(P_{\text{agg},i}, w_i^{\text{cls}})$ and $S_{t2g} = S(v_i^{\text{cls}}, w_i^{\text{cls}})$ are similarity distributions, KL denotes the relative entropy loss function, H represents the entropy regularization term for robustness, and λ is a weighting factor.

3.3.5 Classification Loss. For downstream tasks, such as desire understanding, emotion recognition, and sentiment analysis, we use the standard cross-entropy loss:

$$\mathcal{L}_{cls} = \frac{1}{N}\sum_{i=1}^{N}\text{CrossEntropy}(y_i, \widehat{y}_i) \tag{6}$$

where y is the ground-truth label and $\widehat{y}_i$ is the predicted label.

3.4 Two-stage Training Strategy

To balance semantic alignment with pixel-level reconstruction, we adopt a two-stage training strategy and selectively freeze or unfreeze model components in each stage to steer learning.

3.4.1 Pre-training Stage. The pre-training stage focuses on masked image modeling, giving priority to training the image encoder and the text-guided image decoder. In this stage, we freeze the image-guided text decoder and the MLP, and only update the text encoder, the image encoder, and the text-guided image decoder. This design allows us to extract fine-grained details from high-resolution sub-images while constraining modal consistency at the semantic level. As discussed in Section 3.3.3 and Section 3.3.4, we also consider the sub-image-level global visual representation of reconstructed image tokens and their cross-modal alignment. Therefore, the overall loss used in this stage is:

$$\mathcal{L}_p = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{si}\mathcal{L}_{si} + \lambda_{dc}\mathcal{L}_{dc} + \lambda_{itc}\mathcal{L}_{itc} \tag{7}$$

where λ_{rec}, λ_{si}, λ_{dc} and λ_{itc} are hyperparameter weights that balance the contributions of each loss term.

3.4.2 Fine-tuning Stage. The fine-tuning stage targets downstream tasks such as desire understanding. Its goal is to train the image-guided text decoder to fully leverage both local and global features produced by the image encoder, and to use an MLP to map the fused multimodal representation to task-specific outputs. During fine-tuning we freeze the image encoder and text-guided image decoder, and train the text encoder, the image-guided text decoder, and the MLP. The overall loss for this stage is:

$$\mathcal{L}_f = \lambda_{cls}\mathcal{L}_{cls} + \lambda_{itc}\mathcal{L}_{itc} \tag{8}$$

where λ_{cls} and λ_{itc} are hyperparameter weights. Retaining the ITC term helps preserve cross-modal alignment stability during fine-tuning.

4 Experiments

4.1 Experiment Setup

4.1.1 Dataset. To validate our method, we use the MSED dataset [14], the first multimodal multi-task benchmark for sentiment analysis, emotion recognition, and desire understanding. MSED comprises 9,190 English image–text pairs collected from social media (e.g., Twitter, Getty Images, Flickr). Each sample is annotated with a

Table 1: Statistics of MSED Dataset

Train	Validation	Test	Total
6,127 (66.7%)	1,021 (11.1%)	2,042 (22.2%)	9,190

Table 2: Comparison of SyDES and other SOTA methods on the MSED dataset

Method	Desire Understanding				Emotion Recognition				Sentiment Analysis			
	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
DCNN+AlexNet [14]	59.42	52.02	52.35	-	49.56	42.77	43.76	-	71.02	70.09	70.31	-
DCNN+ResNet [14]	56.34	50.64	52.89	-	62.93	59.12	60.48	-	74.73	74.73	74.64	-
BiLSTM+AlexNet [14]	67.80	68.00	67.67	-	71.17	70.70	70.89	-	78.73	79.22	78.89	-
BERT+AlexNet [14]	80.84	75.50	77.17	-	78.06	78.19	78.10	-	83.22	83.11	83.16	-
Multimodal Transformer [14]	81.92	80.20	80.92	-	81.62	81.61	81.53	-	83.56	83.45	83.50	-
M3GAT [44]	-	-	-	-	82.53	81.51	81.97	-	84.66	85.15	84.85	-
MMTF-DES [1]	**84.23**	82.01	83.11	86.97	84.39	84.64	84.26	84.13	88.27	88.68	88.44	88.44
SyDES (Ours)	84.09	**84.07**	**84.02**	**88.32**	**84.92**	**84.81**	**84.74**	**85.96**	**89.28**	**89.13**	**89.19**	**89.37**
%Gains	-0.20	+2.50	+1.10	+1.60	+0.60	+0.20	+0.60	+2.20	+1.10	+0.50	+0.90	+1.10

Table 3: Comparison between SyDES and different modality baseline models on the MSED dataset

Task	Method	Modality	P	R	F1
Desire Understanding	BERTweet	Text	77.11	81.19	78.86
	ResNet	Image	49.97	49.35	49.20
	SyDES-B	Multimodal	80.77	67.43	72.27
	SyDES	Multimodal	**84.09**	**84.07**	**84.02**
Emotion Recognition	BERTweet	Text	80.99	77.15	78.34
	ResNet	Image	58.74	54.67	56.40
	SyDES-B	Multimodal	81.04	80.66	80.80
	SyDES	Multimodal	**84.92**	**84.81**	**84.74**
Sentiment Analysis	BERTweet	Text	82.25	83.62	82.49
	ResNet	Image	70.85	70.61	70.64
	SyDES-B	Multimodal	89.09	85.58	86.50
	SyDES	Multimodal	**89.28**	**89.13**	**89.19**

sentiment label (*positive, neutral, negative*), an emotion label (*happiness, sad, neutral, disgust, anger, and fear*), and a desire label (*family, romance, vengeance, curiosity, tranquility, social-contact, and none*). The data are split into train, validation, and test set. Detailed statistics are given in Table 1 and the detailed statistics are provided in Appendix B.1.

4.1.2 Evaluation Metrics. All three downstream tasks are classification problems. We therefore report standard classification metrics: Precision (P), Recall (R), Macro-F1-score (F1), and Weighted Accuracy (Acc).

4.1.3 Training Details. All experiments run on NVIDIA V100 with CUDA 11.0 and PyTorch 2.1.2 [24]. We initialize the image encoder and the text encoder from CLIP [26] pre-training weights provided by OpenAI[1], and initialize the text-guided image decoder from pre-training weights[2] in [13]. For complete training hyper-parameter settings are provided in Appendix B.2.

4.1.4 Baseline. To facilitate subsequent ablation analysis and comparison, we use the SyDES architecture trained directly with classification loss without pre-training stage. We denote this variant as **SyDES-B** (the SyDES baseline).

[1]https://github.com/openai/CLIP
[2]https://huggingface.co/laion/mscoco_finetuned_CoCa-ViT-L-14-laion2B-s13B-b90k

4.2 Comparison with different modality baseline models

We compared different modality baseline models on three downstream tasks, including desire understanding, emotion recognition, and sentiment analysis, to evaluate the effectiveness of our proposed SyDES. For the textual modality, we employed **BERTweet** [22] model as the baseline modal since text data are annotated from social media platforms. For the image modality, we used the classic backbone network **ResNet** [8]. The multimodal baseline model, SyDES-B, was included to demonstrate the advantage of multimodal fusion. We also present the results of our proposed SyDES method.

The experiment results for different modality baseline models are shown in Table 3. Analysis of these results leads to three main conclusions: (1) Multimodal models consistently outperform unimodal models across mostly tasks. For instance, in emotion recognition, SyDES-B improved the F1-score by 3.14% gains over BERTweet, indicating the benefit of leveraging multiple modalities in sentiment-related tasks. (2) The unimodal image model (e.g., ResNet) consistently underperformed compared to the unimodal text model (e.g., BERTweet), suggesting limitations in capturing fine-grained visual semantics and underutilization of non-verbal cues. (3) Our proposed SyDES consistently surpassed SyDES-B in all tasks, validating the effectiveness of the non-verbal cues mining mechanism introduced during pre-training.

4.3 Comparison with state-of-the-art methods

The comparative performance of our proposed SyDES on test data against other SOTA methods across all tasks is presented in Table 2. The results indicate that SyDES achieves competitive performance across all three tasks. In terms of the primary metric F1-score, our proposed SyDES surpassed the previous best model, MMTF-DES [1], by 1.1%, 0.6%, and 0.9% gains, respectively. It is worth noting that MMTF-DES relies on integrating multiple multimodal Transformer encoders (e.g., ViLT and VAuLT), which entails considerably higher training costs. In contrast, SyDES extracts non-verbal cues from images effectively while maintaining lower computational overhead, yielding pronounced gains on desire understanding, which depends more heavily on non-verbal cues.

Table 4: Ablation study on the loss functions used in the two training stages

Pre-training				Fine-tuning		Desire Understanding				Emotion Recognition				Sentiment Analysis			
$\mathcal{L}_{\text{rec}}$	$\mathcal{L}_{itc}$	$\mathcal{L}_{si}$	$\mathcal{L}_{dc}$	$\mathcal{L}_{itc}$	$\mathcal{L}_{cls}$	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
×	×	×	×	×	✓	80.77	67.43	72.27	80.51	81.04	80.66	80.80	82.61	89.09	85.58	86.50	87.12
✓	×	×	×	✓	✓	76.17	72.56	74.12	81.15	76.78	75.13	75.82	77.62	80.57	81.19	80.79	80.80
✓	✓	×	×	✓	✓	81.43	81.77	81.44	86.19	84.85	84.95	84.66	85.21	89.26	88.80	88.99	89.18
✓	✓	✓	×	✓	✓	82.28	82.33	82.24	86.78	84.17	81.91	82.88	84.53	87.45	87.66	87.55	87.71
✓	✓	✓	✓	×	✓	78.55	75.89	77.09	83.74	84.85	81.75	83.21	84.43	86.46	85.93	86.16	86.44
✓	✓	✓	✓	✓	✓	**84.09**	**84.07**	**84.02**	**88.32**	**84.92**	84.81	**84.74**	**85.96**	**89.28**	**89.13**	**89.19**	**89.37**

Table 5: Ablation study of proposed components.

Step	Configuration	Desire		Emotion		Sentiment	
		F1	Acc	F1	Acc	F1	Acc
0	*Baseline (SyDES-B)*	72.27	80.51	80.80	82.61	86.50	87.12
1	+ Mixed-scale image strategy	75.49	79.97	78.70	81.49	86.37	86.24
2	+ Reconstruction decoder	81.51	84.96	84.40	85.26	89.03	89.08
3	+ Aggregator	**84.02**	**88.32**	**84.74**	**85.96**	**89.19**	**89.37**

We further observe that sentiment analysis generally yields higher performance among the three tasks. A possible reason for this difference in performance may be attributed to the nature of the three tasks. Sentiment analysis aim to identify the overall emotion or opinion expressed in an image-text pair. In contrast, desire understanding and emotion recognition require fine-grained detection of specific signals such as a person's gestures or facial expressions that are inherently embedded in images. For example, there exists the exaggerated motion and frightened expression of the man in Figure 1 (b). These subtle cues are inherently more challenging to capture. Therefore, sentiment analysis may be an easier and more straightforward task for model, while desire understanding and emotion recognition may be more complicated and nuanced. Our proposed SyDES enhances local detail extraction, resulting in particularly notable gains in desire understanding, but its performance remains slightly below that of sentiment analysis. This suggests there is still room for improvement and underscores the need for further research into desire understanding and emotion recognition.

4.4 Ablation Studies

4.4.1 Loss Functions. We adopt a two-stage training strategy to facilitate cross-modal fusion between textual features and both local and global visual features. As summarized in Table 4, we systematically evaluate the impact of different loss combinations on the three sub-tasks. The first row corresponds to the model fin-tuned directly without pre-training stage (i.e., SyDES-B as described in Section 4.1.4). During pre-training, four loss functions are used, including $\mathcal{L}_{rec}$, $\mathcal{L}_{si}$, $\mathcal{L}_{dc}$, and $\mathcal{L}_{itc}$. Results show that using only $\mathcal{L}_{rec}$ for masked image modeling leads to performance even worse than SyDES-B. This indicates that reconstructing high-resolution sub-images alone may introduce a semantic mismatch with the low-resolution image due to inadequate modal alignment, resulting in modal inconsistency that harms fine-tuning. Gradually incorporating $\mathcal{L}_{itc}$, $\mathcal{L}_{si}$, and $\mathcal{L}_{dc}$ consistently improves performance across all tasks. For example, in desire understanding, the F1-score increases to 81.44% with $\mathcal{L}_{itc}$, to 82.24% with $\mathcal{L}_{si}$, and finally to 84.02% with $\mathcal{L}_{dc}$. This suggests that cross-modal alignment and semantic/consistency constraints are critical to bridging the gap between reconstructed sub-images and the global image semantics. More detailed result can be found in Appendix C.2.

During fine-tuning, two loss functions are used, including $\mathcal{L}_{itc}$ and $\mathcal{L}_{cls}$. We observe that that $\mathcal{L}_{itc}$ is important to preserve pre-training gains. Removing $\mathcal{L}_{itc}$ during fine-tuning causes the F1-scores to drop from 84.02%, 84.74%, and 89.19% to 77.09%, 83.21%, and 86.16%, respectively. In conclusion, using all proposed loss functions yields the best performance across desire understanding, emotion recognition, and sentiment analysis, confirming the complementary effects of the loss terms.

4.4.2 Contribution of proposed modules. As shown in Table 5, the results demonstrate that the complete model achieves the best performance across three tasks, with the F1 score for desire understanding significantly increasing from 72.27% to 84.0%. It is worth noting that mixed-scale image strategy alone leads to a performance drop in desire understanding, indicating that simple modality dose not fully exploit fine-gained semantic alignment. Introducing the reconstruction decoder substantially improves desire understanding, demonstrating that fine-grained reconstruction effectively enhances the model's sensitivity to local semantics. Moreover, incorporating the aggregator along with consistency losses further improves all metrics, highlighting the importance of global-local semantic coordination for effective learning.

4.4.3 Ratio of Masked Tokens. As shown in Table 6, we investigate the impact by setting different ratios of masked tokens. Specifically, 0.25, 0.50, 0.75, and 0.90 are tested. Experimental results demonstrate that a masking ratio of 75% yields the best average performance across most tasks. Although a ratio of 25% achieves slightly better results in emotion recognition, the 75% ratio is overall more suitable considering its substantially lower computational cost while maintain competitive performance.

4.5 Visualization

4.5.1 Cross-Modal Attention Heatmap Analysis. To intuitively illustrate the perception ability of our proposed SyDES toward textual and image modalities, we visualized the attention localization map for the last layer in the image encoder. As depicted in Figure 4, our SyDES focuses more accurately on fine-grained visual regions corresponding to informative words in the text. For example, in the

Table 6: Ablation study on the ratio of masked tokens

Ratio of Masked Tokens	Desire Understanding				Emotion Recognition				Sentiment Analysis			
	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
0.25	83.13	81.80	82.23	86.82	**85.94**	**85.30**	**85.55**	85.94	88.90	89.05	88.87	88.83
0.50	81.87	81.69	81.69	86.48	84.36	84.35	84.19	85.06	87.64	88.18	87.88	88.06
0.75	**84.09**	**84.07**	**84.02**	**88.32**	84.92	84.81	84.74	**85.96**	**89.28**	**89.13**	**89.19**	**89.37**
0.90	81.31	82.82	82.02	86.48	84.32	83.31	83.78	85.41	88.54	87.93	88.17	88.39

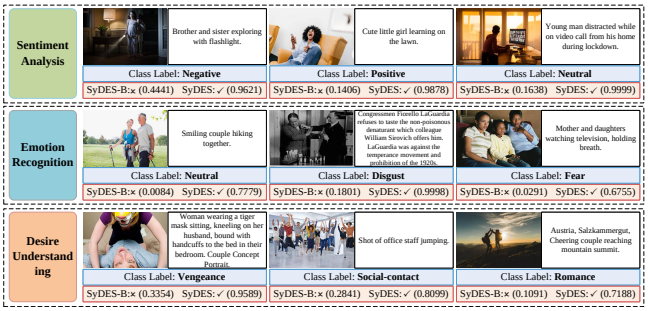


Figure 3: Performance analysis of our proposed SyDES on desire understanding, emotion recognition, and sentiment analysis tasks. × indicates correct classification and ✓ represents misclassification.

case of Image 1, SyDES shows concentrated attention on regions related to "***boy***" and "***bike***", whereas SyDES-B neglects these essential details. Similarity, in the case of Image 3, SyDES-B perceives the concept of "***family***" vaguely, while our proposed SyDES clearly identifies the five-person "***family***" and the "***beach***" scene.

4.5.2 Reconstructed Image Visualization. We reconstructed and stitched the high-resolution sub-images, and visualized the masked image patches to inspect the practical effect in Figure 5. Each sample consists of the original image, the randomly masked image, and the reconstructed image by the text-guided image decoder. The results show that, even with complex image content (e.g., *people, gesture, lighting, and natural scenes* in the example 1), our proposed SyDES can reconstruct contextual content reasonably well, owing to masked image modeling using high-resolution sub-images and text-guided decoding.

4.5.3 Performance Analysis. To better understand the advantage of our proposed SyDES, we selected representative examples from the MSED dataset across three sub-tasks. As illustrated in Figure 3, we compare the predictions of SyDES-B and our proposed SyDES, including example stimuli, ground-truth labels, and predicted probabilities. Across all three tasks, our proposed SyDES produces correct predictions with high confidence. For instance, in Example 1 of the sentiment analysis task, the image caption is "***Brother and sister exploring with flashlight.***", it provides limited information. But our proposed SyDES achieved a high confidence score of 96.21% by effectively leveraging visual cues.

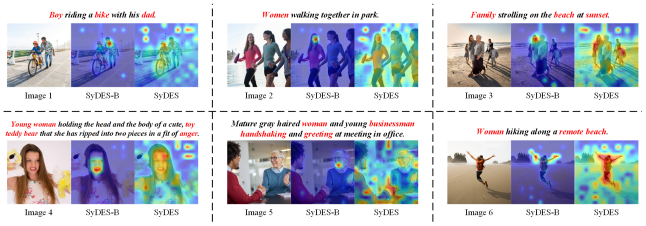


Figure 4: Qualitative analysis of attention localization map in our proposed SyDES. We visualize the attention localization map from the last layer in the image encoder, reflecting the model's cross-modal perception of text.

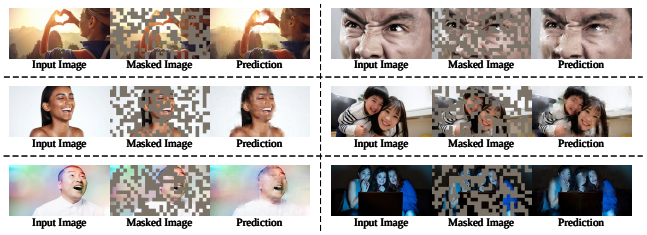


Figure 5: Visualization of reconstructed images using our proposed SyDES.

5 Conclusion

To address the under-exploration of desire understanding and the inadequate use of non-verbal cues, we propose a symmetric bidirectional multimodal learning framework for desire, emotion, and sentiment recognition. The framework employs a mixed-scale image strategy, using masked image modeling to boosting extract fine-grained local features from high-resolution sub-images and maintaining global context from low-resolution images. Furthermore, its symmetrical cross-modal decoders, including a text-guided image decoder and an image-guided text decoder, enable bidirectional interaction that refines both visual and textual representations. To harmonize these objectives, a set of dedicated loss functions are used, facilitating consistent learning across modalities. Extensive experiments on the MSED dataset demonstrate the effectiveness of leveraging non-verbal cues.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62272188).

References

[1] Abdul Aziz, Nihad Karim Chowdhury, Muhammad Ashad Kabir, Abu Nowshed Chy, and Md. Jawad Siddique. 2025. MMTF-DES: A fusion of multimodal transformer models for desire, emotion, and sentiment analysis of social media data. *Neurocomputing* 623 (2025), 129376. https://api.semanticscholar.org/CorpusID:275561806

[2] Stephanie Cacioppo, Francesco Bianchi-Demicheli, Chris Frum, James G Pfaus, and James W Lewis. 2012. The common neural bases between sexual desire and love: a multilevel kernel density fMRI analysis. *The journal of sexual medicine* 9, 4 (2012), 1048–1054.

[3] Dushyant Singh Chauhan, SR Dhanush, Asif Ekbal, and Pushpak Bhattacharyya. 2020. All-in-one: A deep attentive multi-task learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes. In *Proceedings of the 1st conference of the Asia-Pacific chapter of the association for computational linguistics and the 10th international joint conference on natural language processing*. 281–290.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv* abs/2010.11929 (2020). https://api.semanticscholar.org/CorpusID:225039882

[5] Pablo Gamallo and Marcos Garcia. 2014. Citius: A naive-bayes strategy for sentiment analysis on english tweets. In *Proceedings of the 8th international Workshop on Semantic Evaluation (SemEval 2014)*. 171–175.

[6] Wenya Guo, Ying Zhang, Xiangrui Cai, L. Meng, Jufeng Yang, and Xiaojie Yuan. 2021. LD-MAN: Layout-Driven Multimodal Attention Network for Online News Sentiment Recognition. *IEEE Transactions on Multimedia* 23 (2021), 1785–1798. https://api.semanticscholar.org/CorpusID:226742001

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. 2021. Masked Autoencoders Are Scalable Vision Learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 15979–15988. https://api.semanticscholar.org/CorpusID:243985980

[8] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 770–778. https://api.semanticscholar.org/CorpusID:206594692

[9] Wilhelm Hofmann and Loran F Nordgren. 2015. *The psychology of desire*. Guilford Publications.

[10] Sabrina Hoppe, Tobias Loetscher, Stephanie Morey, and Andreas Bulling. 2015. Recognition of curiosity using eye movement analysis. *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers* (2015). https://api.semanticscholar.org/CorpusID:15967389

[11] Anthony Hu and Seth Flaxman. 2018. Multimodal Sentiment Analysis To Explore the Structure of Emotions. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018). https://api.semanticscholar.org/CorpusID:44075392

[12] Feiran Huang, Kaimin Wei, Jian Weng, and Zhoujun Li. 2020. Attention-Based Modality-Gated Networks for Image-Text Sentiment Analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16 (2020), 1 – 19. https://api.semanticscholar.org/CorpusID:218517893

[13] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. doi:10.5281/zenodo.5143773 If you use this software, please cite it as below..

[14] Ao Jia, Yu He, Yazhou Zhang, Sagar Uprety, Dawei Song, and Christina Lioma. 2022. Beyond Emotion: A Multi-Modal Dataset for Human Desire Understanding. In *North American Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:250391079

[15] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:9672033

[16] Hoai-Duy Le, Gueesang Lee, Soo hyung Kim, Seung won Kim, and Hyung-Jeong Yang. 2023. Multi-Label Multimodal Emotion Recognition With Transformer-Based Fusion and Emotion-Level Representation Learning. *IEEE Access* 11 (2023), 14742–14751. https://api.semanticscholar.org/CorpusID:256944875

[17] Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. CLMLF: A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection. *ArXiv* abs/2204.05515 (2022). https://api.semanticscholar.org/CorpusID:248119031

[18] Angelica Lim, Tetsuya Ogata, and Hiroshi G Okuno. 2012. The desire model: Cross-modal emotion analysis and expression for robots. *Information Processing Society of Japan* 5, 4 (2012).

[19] Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Clarence Yan, and Tao Mei. 2019. Social Relation Recognition From Videos via Multi-Scale Spatial-Temporal Reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 3561–3569. https://api.semanticscholar.org/CorpusID:198118474

[20] Ilya Loshchilov and Frank Hutter. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*. https://api.semanticscholar.org/CorpusID:53592270

[21] Shahla Nemati, Reza Rohani, Mohammad Ehsan Basiri, Moloud Abdar, Neil Yuwen Yen, and Vladimir Makarenkov. 2019. A Hybrid Latent Space Data Fusion Method for Multimodal Emotion Recognition. *IEEE Access* 7 (2019), 172948–172964. https://api.semanticscholar.org/CorpusID:209320752

[22] Dat Quoc Nguyen, Thanh Vu, and Anh Gia-Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:218719869

[23] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070* (2002).

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *ArXiv* abs/1912.01703 (2019). https://api.semanticscholar.org/CorpusID:202786778

[25] Paul Portner and Aynat Rubinstein. 2020. Desire, belief, and semantic composition: variation in mood selection with desire predicates. *Natural Language Semantics* 28 (2020), 343 – 393. https://api.semanticscholar.org/CorpusID:226352313

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:231591445

[27] Hiranmayi Ranganathan, Shayok Chakraborty, and Sethuraman Panchanathan. 2016. Multimodal emotion recognition using deep learning architectures. *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2016), 1–9. https://api.semanticscholar.org/CorpusID:6182290

[28] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The Graph Neural Network Model. *IEEE Transactions on Neural Networks* 20 (2009), 61–80. https://api.semanticscholar.org/CorpusID:206756462

[29] Nicola S. Schutte and John M. Malouff. 2019. A Meta-Analysis of the Relationship between Curiosity and Creativity. *The Journal of Creative Behavior* (2019). https://api.semanticscholar.org/CorpusID:199157255

[30] Mrityunjay Singh, Amit Kumar Jakhar, and Shivam Pandey. 2021. Sentiment analysis on the impact of coronavirus in social life using the BERT model. *Social Network Analysis and Mining* 11 (2021). https://api.semanticscholar.org/CorpusID:232293517

[31] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37, 2 (2011), 267–307.

[32] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *ArXiv* abs/1503.00075 (2015). https://api.semanticscholar.org/CorpusID:3033526

[33] Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. 2024. WisdoM: Improving Multimodal Sentiment Analysis by Fusing Contextual World Knowledge. *Proceedings of the 32nd ACM International Conference on Multimedia* (2024). https://api.semanticscholar.org/CorpusID:266977237

[34] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016. Attention-based LSTM for Aspect-level Sentiment Classification. In *Conference on Empirical Methods in Natural Language Processing*. https://api.semanticscholar.org/CorpusID:18993998

[35] Nan Xu. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)* (2017), 152–154. https://api.semanticscholar.org/CorpusID:20067030

[36] Nan Xu and Wenji Mao. 2017. MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (2017). https://api.semanticscholar.org/CorpusID:29030535

[37] Nan Xu, Wenji Mao, and Guandan Chen. 2018. A Co-Memory Network for Multimodal Sentiment Analysis. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (2018). https://api.semanticscholar.org/CorpusID:195351351

[38] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2020. Image-Text Multimodal Emotion Classification via Multi-View Attentional Network. *IEEE Transactions on Multimedia* 23 (2020), 4014–4026. https://api.semanticscholar.org/CorpusID:229272644

[39] Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. In *Annual Meeting of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:236460184

[40] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard H. Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *North American Chapter of the Association for Computational Linguistics*. https://api.semanticscholar.org/CorpusID:6857205

[41] Özerk Yavuz, Adem Karahoca, and Dilek Karahoca. 2019. A data mining approach for desire and intention to participate in virtual communities. *International Journal of Electrical and Computer Engineering (IJECE)* (2019). https://api.semanticscholar.org/CorpusID:208980057

[42] Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. 2016. Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia. *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (2016). https://api.semanticscholar.org/CorpusID:7928793

[43] Yong Zhang, Cheng Cheng, and YiDie Zhang. 2022. Multimodal emotion recognition based on manifold learning and convolution neural network. *Multimedia Tools and Applications* 81 (2022), 33253 – 33268. https://api.semanticscholar.org/CorpusID:248252616

[44] Yazhou Zhang, Ao Jia, Bo Wang, Peng Zhang, Dongming Zhao, Pu Li, Yuexian Hou, Xiaojia Jin, Dawei Song, and Jing Qin. 2023. M3GAT: A Multi-modal, Multi-task Interactive Graph Attention Network for Conversational Sentiment Analysis and Emotion Recognition. *ACM Transactions on Information Systems* 42 (2023), 1 – 32. https://api.semanticscholar.org/CorpusID:258788073

[45] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, and Xiao Xiao. 2023. Multimodal Emotion Classification With Multi-Level Semantic Reasoning Network. *IEEE Transactions on Multimedia* 25 (2023), 6868–6880. https://api.semanticscholar.org/CorpusID:253313650

A Limitation

Although we thoroughly explore non-verbal cues in images through a masked image modeling and our proposed method demonstrated effectiveness on the MSED dataset, several limitations remain:

(1) Capturing fine-grained local features is still insufficient. As shown in Section 4.5.2, reconstructions of small but semantically important region such as faces are limited, indicating that there is still room for improvement in perceiving and reconstructing fine-grained regions.
(2) The exploitation of textual information could be further enhanced. Textual descriptions often obtain abundant aspect-level expressions (e.g., opinion words). Effectively leveraging such terms, integrating and aligning them with visual information represents a promising direction for future research.
(3) The current method focuses only on image-text pairs. Real-world multimodal data involve additional modalities (e.g., video and audio), which provide richer non-verbal cues. Extending our method to incorporate more modalities while balancing computational cost and performance constitutes an important direction for further investigation.

Table 7: Data statistics of MSED dataset

Task	Label	Train	Validation	Test
Desire Understanding	Vengeance	277	39	75
	Curiosity	634	118	213
	Social-contact	437	59	138
	Family	873	152	288
	Tranquility	245	39	87
	Romance	692	101	210
	None	2,969	507	1,031
Emotion Recognition	Happiness	2,524	419	860
	Sad	666	102	186
	Neutral	1,664	294	569
	Disgust	251	44	80
	Anger	523	78	172
	Fear	499	84	175
Sentiment Analysis	Positive	2,524	419	860
	Neutral	1,664	294	569
	Negative	1,939	308	613

B Dataset and Training Details

B.1 Dataset Statistics

The complete, per-class detailed statistics are provided in Table 7.

B.2 Training Hyperparameters

The complete training and implementation details for both pre-training and fine-tuning stages are summarized in Table 8.

C Detailed Experimental Results

C.1 Contributions of Different Modalities

As shown in Table 9, we conduct an experiment to analyze the contribution of non-verbal cues. Our method achieves a significant improvement under the image-only setting. This indicates that our method can effectively utilize non-verbal cues from images. In contrast, performance gains are limited or even degrade under the text-only setting, further confirming that our model focuses more on visual information. In the multimodal setting, our method consistently outperforms the baseline, indicating that it enhances visual signal while preserving textual understanding. These results demonstrate that the effectiveness of our proposed method in leveraging non-verbal cues from images.

C.2 Contributions of Loss Functions

To provide a more comprehensive contribution analysis of each loss function in our proposed framework, we conduct extensive ablation studies as shown in Table 10. The experiments systematically investigate the necessity and effectiveness of each loss function in both pre-training and fine-tuning stages.

Table 8: Pretraining and fine-tuning hyperparameters

Hyperparameter	Pre-training	Fine-tuning
Configuration		
Batch Size	64	64
Epochs	50	50
Optimizer	AdamW [20]	
LR Schedule	Cosine Decay	
Weight Decay	0.01	0.01
Warmup Ratio	0.15	0.10
Mask Ratio	0.75	0.00
Learning Rates		
Image Encoder	5×10^{-6}	-
Text Encoder	5×10^{-5}	1×10^{-4}
Image Decoder	1×10^{-4}	-
Text Decoder	-	2×10^{-4}
MLP Classifier	-	1×10^{-4}
Loss Weights		
λ_{rec}	1.0	-
λ_{si}	0.5	-
λ_{dc}	0.025	-
λ_{itc}	0.5	0.4
λ_{cls}	-	1.0

C.2.1 Experimental Design. Our ablation study includes the following configurations:

- **Baseline (Row 1)**: Training with only classification loss $\mathcal{L}_{\text{cls}}$ in fine-tuning stage, without any pre-training.
- **Individual Loss Analysis (Rows 2-5)**: We first examine individual components separately:
  - Only $\mathcal{L}_{\text{rec}}$ in pre-training (Row 2-3)
  - Only $\mathcal{L}_{\text{itc}}$ in pre-training (Row 4-5)
- **Loss Combinations (Rows 6-8)**: We investigate the interaction between different losses:
  - $\mathcal{L}_{\text{rec}}$ + $\mathcal{L}_{\text{itc}}$ combination (Row 6)
  - $\mathcal{L}_{\text{rec}}$ + $\mathcal{L}_{\text{si}}$ + $\mathcal{L}_{\text{dc}}$ combination (Row 7)
  - $\mathcal{L}_{\text{rec}}$ + $\mathcal{L}_{\text{itc}}$ + $\mathcal{L}_{\text{si}}$ combination (Row 8)
- **Stage-specific Ablations (Row 9)**: We examine the importance of $\mathcal{L}_{\text{itc}}$ in fine-tuning stage.
- **Complete Model (Row 10)**: Our complete framework with all losses.

C.2.2 Key Findings. From the comprehensive results, we observe several important patterns:

Effectiveness of $\mathcal{L}_{\text{itc}}$. The image-text contrastive loss plays a crucial role in both stages. When $\mathcal{L}_{\text{itc}}$ is removed from pre-training (Row 7), performance drops significantly across all tasks. Similarly, removing $\mathcal{L}_{\text{itc}}$ from fine-tuning (Row 9) also leads to notable performance degradation, confirming its necessity in both stages.

Synergy between Losses. The combination of $\mathcal{L}_{\text{rec}}$ and $\mathcal{L}_{\text{itc}}$ (Row 6) shows substantial improvement over using either component alone (Rows 2-5), demonstrating their complementary nature. Adding $\mathcal{L}_{\text{si}}$ (Row 8) further improves Desire Understanding performance, validating the importance of structural information.

Progressive Improvement. Performance improves progressively as we add more constrained losses, with the complete model achieving the best results across all metrics.

Cross-Task Consistency. The ablation patterns are consistent across all three tasks, suggesting that our loss components provide general benefits rather than task-specific optimizations.

Table 9: Complete experimental results on the contributions of different Modalities. We report Precision (P), Recall (R), F1-score (F1), and Accuracy (Acc). Under the *Image-Only* setting, the input text is an empty string; under the *Text-Only* setting, the input image is an all-one tensor. The baseline is SyDES-B.

Methods	Modal		Desire Understanding				Emotion Recognition				Sentiment Analysis			
	Text	Image	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
SyDES-B	×	✓	49.07	28.92	25.16	50.20	69.15	63.25	62.52	72.92	78.71	78.71	78.02	79.04
	✓	×	40.32	29.05	29.00	58.08	68.01	65.06	62.35	62.14	77.86	74.01	72.89	71.79
	✓	✓	80.77	67.43	72.27	80.51	81.04	80.66	80.80	82.61	89.09	85.58	86.50	87.12
SyDES	×	✓	75.61	63.35	65.62	77.91	72.47	69.74	70.71	76.10	83.81	81.07	82.02	82.52
	✓	×	43.96	35.82	37.24	58.67	65.08	48.54	49.53	62.49	78.30	74.02	72.46	71.35
	✓	✓	84.09	84.07	84.02	88.32	84.92	84.81	84.74	85.96	89.28	89.13	89.19	89.37

Table 10: Comprehensive ablation study on the loss functions in the two-stage training

Pre-training				Fine-tuning		Desire Understanding				Emotion Recognition				Sentiment Analysis			
$\mathcal{L}_{rec}$	$\mathcal{L}_{itc}$	$\mathcal{L}_{si}$	$\mathcal{L}_{dc}$	$\mathcal{L}_{itc}$	$\mathcal{L}_{cls}$	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
×	×	×	×	×	✓	80.77	67.43	72.27	80.51	81.04	80.66	80.80	82.61	89.09	85.58	86.50	87.12
✓	×	×	×	×	✓	73.01	72.41	72.63	79.33	75.83	73.55	73.99	75.86	80.00	80.90	80.33	80.51
✓	×	×	×	✓	✓	76.17	72.56	74.12	81.15	76.78	75.13	75.82	77.62	80.57	81.19	80.79	80.80
×	✓	×	×	×	✓	78.58	75.97	76.76	82.71	82.02	81.65	81.82	83.38	86.54	86.90	86.70	86.86
×	✓	×	×	✓	✓	80.26	81.95	80.96	84.19	83.09	82.15	82.57	84.43	86.41	86.89	86.61	86.64
✓	✓	×	×	✓	✓	81.43	81.77	81.44	86.19	84.85	84.95	84.66	85.21	89.26	88.80	88.99	89.18
✓	×	✓	✓	✓	✓	75.45	73.52	74.39	80.31	76.49	74.05	75.07	77.23	80.51	80.95	80.71	80.85
✓	✓	✓	×	✓	✓	82.28	82.33	82.24	86.78	84.17	81.91	82.88	84.53	87.45	87.66	87.55	87.71
✓	✓	✓	✓	×	✓	78.55	75.89	77.09	83.74	84.85	81.75	83.21	84.43	86.46	85.93	86.16	86.44
✓	✓	✓	✓	✓	✓	**84.09**	**84.07**	**84.02**	**88.32**	**84.92**	**84.81**	**84.74**	**85.96**	**89.28**	**89.13**	**89.19**	**89.37**