Synergy of Sight and Semantics: Visual Intention Understanding with CLIP

Qu Yang[1], Mang Ye[1*], and Dacheng Tao[2]

[1] National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, China
[2] Nanyang Technological University, Singapore
{yangqu,yemang}@whu.edu.cn dacheng.tao@gmail.com

Abstract. Multi-label Intention Understanding (MIU) for images is a critical yet challenging domain, primarily due to the ambiguity of intentions leading to a resource-intensive annotation process. Current leading approaches are held back by the limited amount of labeled data. To mitigate the scarcity of annotated data, we leverage the Contrastive Language-Image Pre-training (CLIP) model, renowned for its wealth knowledge in textual and visual modalities. We introduce a novel framework, **Intention Understanding with CLIP** (IntCLIP), which utilizes a dual-branch approach. This framework exploits the 'Sight'-oriented knowledge inherent in CLIP to augment 'Semantic'-centric MIU tasks. Additionally, we propose **Hierarchical Class Integration** to effectively manage the complex layered label structure, aligning it with CLIP's nuanced sentence feature extraction capabilities. Our **Sight-assisted Aggregation** further refines this model by infusing the semantic feature map with essential visual cues, thereby enhancing the intention understanding ability. Through extensive experiments conducted on the standard MIU benchmark and other subjective tasks such as Image Emotion Recognition, IntCLIP clearly demonstrates superiority over current state-of-the-art techniques. Code is available at `https://github.com/yan9qu/IntCLIP`.

Keywords: Visual Intention Understanding · CLIP

1 Introduction

In the digital age, images on social media platforms carry a multitude of implicit intentions, transcending their explicit visual content. These images aim to persuade, celebrate, inform, and more, highlighting the importance of understanding their underlying intent in areas such as mental health monitoring [26], combating misinformation [37], and others. Despite advancements in Multi-label Intention Understanding (MIU), the field is still hampered by a lack of labeled data, exacerbated by the inherently ambiguous nature of intention. As depicted in Fig. 1 (b), an image often embodies the uncertainty and diversity of the intentions

* Corresponding Author

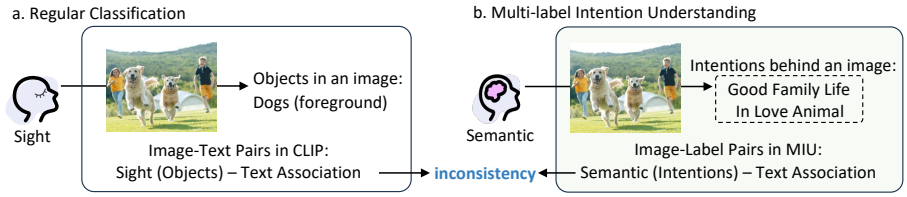

Fig. 1: Inconsistency between sight focus in CLIP and semantic requirements in MIU task. (a) Regular image-based classification tasks, as conceptualized in CLIP, rely on sight—the discernible visual elements such as objects within images—leading to straightforward image-text associations. **(b)** Conversely, Multi-label Intention Understanding (MIU) gravitates towards semantic information, encapsulating complex, often subjective intentions that are not immediately apparent, thus necessitating an extensive and costly labeling process.

behind it. In contrast, fields like standard image classification have flourished by leveraging pre-trained knowledge, such as CLIP [28], drawing on extensive image-text pair collections. This effectively mitigates the limitations posed by scarce labeled data, as shown in Fig. 1 (a). Such knowledge has markedly enhanced feature representations, a potential that we aim to channel into MIU.

However, the knowledge inherent in CLIP is predominantly sight-centric, focusing on observable elements such as the shapes, colors, and spatial arrangements of objects within images. This emphasis creates a notable gap in addressing the semantic insights that are crucial for MIU, which encompass recognizing context, cultural significance, and emotional resonance. These aspects, inferred from an image, extend beyond its literal representation to include the contextual meaning behind the visual content. Current adaptation strategies, particularly prompt learning as explored in [38, 57], tend to overemphasize sight knowledge, thus complicating the integration with semantic intention understanding. Conversely, allowing complete adaptability of the network during training risks compromising the objective-oriented (sight) capabilities learned during pre-training, a concern we elaborate on in our ablation study in Sec. 4.3. The core issue arises from the significant divergence between semantic intent-oriented tasks and objective sight pre-training, highlighting the need for a balanced approach that effectively synergizes both forms of knowledge.

To address this challenge, we introduce the Intention Understanding with CLIP (IntCLIP) framework. IntCLIP is designed to capitalize on sight knowledge while pinpointing semantic intent cues. It features a dual-branch architecture: one branch maintains the original CLIP's immutable image encoding parameters, while the other semantically adaptable branch evolves to focus on intent cues during training, as illustrated in Fig. 2. This configuration facilitates the capture of both sight and semantic information, resulting in a comprehensive image feature representation. Moreover, we introduce Hierarchical Class Integration to effectively leverage multi-layer labels, suitable for sentence-level

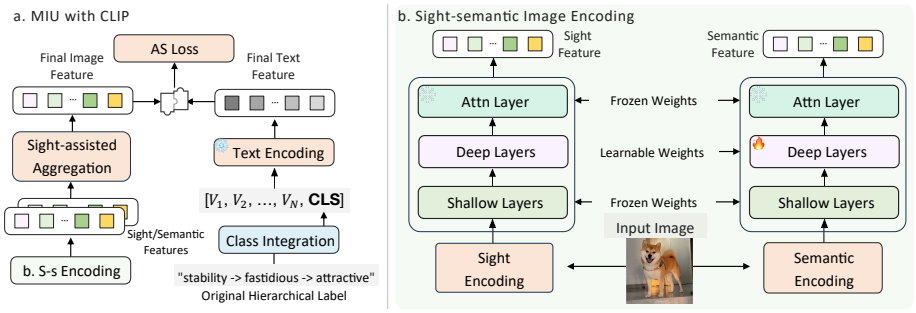

Fig. 2: Overview of proposed Intention Understanding with CLIP (IntCLIP). (a) Demonstrates the MIU with CLIP process, incorporating Sight-assisted Aggregation to fuse sight and semantic features with text encoding, followed by Asymmetric Loss (AS Loss) optimization to align final image and text features. The Hierarchical Class Integration extracts abundant class information from hierarchical labels. **(b)** Depicts the Sight-semantic Image Encoding strategy, where the CLIP-initialized encoder is dedicated to capturing sight-related features, and the semantic branch, through partially learnable deep layers, evolves to accommodate semantic information, thereby preserving both objective visual details and subjective semantic insights.

nature embedded in CLIP. Additionally, our Sight-assisted Aggregation method enhances the sight feature map, aiding in the discernment of critical regions for intention understanding.

We summarize our contributions as follows:

- We are the first to adapt CLIP's sight knowledge to the domain of MIU, significantly enhancing its reasoning capabilities.
- We introduce the IntCLIP framework, along with Hierarchical Class Integration, and the Sight-assisted Aggregation to effectively facilitate this adaptation. Their synergy is instrumental in guiding the extraction of intent cues and pinpointing key intention-related areas.
- Our approach has been empirically validated through experiments on the MIU benchmark and in other subjective tasks such as Image Emotion Recognition (IER), demonstrating marked superiority over existing methods.

2 Related works

2.1 Understanding Human Intention

Understanding human intentions has garnered significant interest in recent years, driven by the improvements of deep learning [2, 4, 12–14, 34, 45–47, 49, 51] and its considerable application potential in practical scenarios. Initial research efforts focused on inferring intentions from facial expressions, bodily cues, and the context of the scene, as explored in studies like [18, 19]. Recently, the research emphasis has shifted towards predicting the underlying motivations of human actions, as evidenced by works such as [20, 40, 43].

In contemporary research, the domain of intention understanding spans diverse fields, including transportation [23, 58], social media [9, 36], and advertising [42]. Pioneering this domain, Jia *et al.* [17] introduced Intentonomy, a fine-grained image-based intention dataset. Their work systematically explored the relationship among objects, contexts, and intentions in images. Concurrently, Shi *et al.* [32, 33, 50] utilized semantic information as guidance for more accurate intention modeling. Similarly, Wang *et al.* [41] proposed a prototype-driven model to address challenges such as intra-class variability and inter-class confusion. Despite their commendable achievements, a persistent semantic mismatch between visual data and textual intention labels remains a challenge. We aim to bridge this gap by incorporating a comprehensive visual-language framework.

2.2 Leveraging Visual-Language Models

Vision-Language models, especially those based on contrastive learning, have demonstrated remarkable abilities to bridge visual and textual data [15, 28]. The pioneering CLIP model [28], trained on an extensive corpus of 400 million image-text pairs, has established new benchmarks in transfer learning capabilities across various classification tasks. This breakthrough has inspired subsequent research [8, 44, 48, 55], investigating optimal training methodologies for a range of downstream applications.

Rather than fine-tuning entire models – a strategy that may disrupt pre-established alignments as highlighted in [6, 10] – recent trends favor a prompt-based paradigm. This approach reimagines classification challenges akin to Masked Language Modeling (MLM), as described in [21,22,35,38]. Pioneering this initiative, Zhou *et al.* [56,57] were among the first to apply CLIP to classification tasks using this paradigm. In parallel, Huang *et al.* [11] introduced a novel method of generating pseudo-labels for images to refine prompts in an unsupervised manner. Sun *et al.* [38] further advanced this approach by presenting dual (both positive and negative) prompts to efficiently transfer alignment knowledge to multi-label classification tasks. However, as discussed in Sec. 1, both all-learning and prompt-learning paradigms have their inherent limitations. Despite these advancements, finding a balance between utilizing pre-trained sight knowledge and adapting it for semantic intent-based tasks remains a challenge. Our Int-CLIP methodology offers an innovative solution to this issue, seeking a balance for intention-centric applications.

3 Proposed Method

3.1 Problem Definition

Multi-label intention understanding can be formally described as follows: Let M represent the set of intention categories which characterize objects or attributes within images. Given a training image with intentions denoted by I, each intention category $m \in M$ is assigned a label y_m, where $y_m = 1$ indicates presence and $y_m = -1$ indicates absence. During the inference phase, the objective is to predict the relevant categories for a new image.

3.2 Revisiting CLIP in Classification

The Contrastive Language-Image Pre-training (CLIP) technique, as introduced by Radford *et al.* [28], demonstrates significant potential in capturing and representing visual cues. Central to CLIP's architecture are dual encoders: one specifically designed for processing images and another for text (refer to Fig. 2 (a) for details). The image encoder is highly versatile, compatible with diverse architectures such as ResNet [10] and Vision Transformer (ViT) [7], effectively translating images into features. Conversely, the text encoder employs a Transformer-based design [29, 39], which converts sequences of word tokens into representations.

For intention fine-tuning, we draw inspiration from Sun et al. [38], adopting a shared embedding space for both visual and textual modalities, given its demonstrated efficacy in multi-label scenarios. Specifically, for a given image-text mini-batch, our goal is to maximize the similarity between correctly paired image-text data, while simultaneously minimizing the similarity with incorrectly paired text. Let x represent the image features, and let $\{w_i\}_{i=1}^{K}$ denote the series of weight vectors from the text encoder, where each vector corresponds to one of intention categories K. Each weight vector w_i is derived from a structured prompt, typically phrased as "a photo depicting a {intention class}" with the placeholder substituted by the corresponding category name. We introduce a learnable prompt "X" to replace the traditional phrasing. For classification, we utilize a combination of softmax computations and cosine similarities to determine probable intention category for each image.

3.3 Infusing Sight Knowledge to Semantic Task

Hierarchical Class Integration via Large Language Models: Contemporary datasets often feature multi-level categorization systems, illustrating label taxonomies in a hierarchical manner. This layered labeling offers deep insights into the complex interrelations among labels, particularly in subjective domains such as multi-label intention understanding (MIU). The extraction of hierarchical labels, spanning from coarse to fine-grained, poses a unique challenge in accurately capturing the layered semantics inherent in multi-label datasets. A basic approach might concatenate labels across various levels using delimiters, represented as $a_c = \{a_1, \ldots, a_n\}$, where a_i represents the label at the i-th level, and a_c constitutes the consolidated label. However, this method could result in disjointed and redundant information.

In contrast to the basic concatenation method, Shi *et al.* [33] proposed Hierarchical Label Embedding and Grouping (HLEG), which utilizes a hierarchical transformer structure to coherently model multi-tiered labels. However, the application of hierarchical transformers may lead to issues of information and parameter redundancy. Specifically, hierarchical labels can exhibit overlapping information (for example, the 'others' category, as illustrated in Fig. 3), and hierarchical networks might introduce excessive parameters when modeling redundant content.

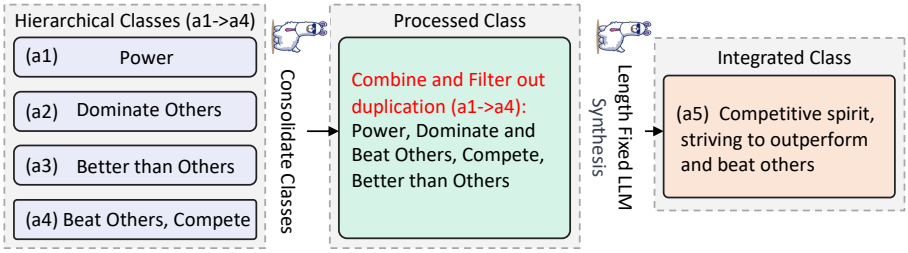

Fig. 3: Hierarchical Class Integration via LLM. (1) Initial hierarchical classes (a1-a4) are consolidated to eliminate redundancy, creating a processed class by filtering out duplicate concepts. **(2)** The processed class is succinctly reformulated into a fixed-length, semantically enriched descriptor via Large Language Model (LLM) synthesis. The integrated sentence-level classes align closely with the inherited nature in CLIP.

To circumvent these challenges while harnessing the adaptable nature of the CLIP text encoder, we propose a two-stage hierarchical label integration strategy. As depicted in Fig. 3, our first stage involves curating labels to eliminate redundant information and establish a set of collated labels via Large Language Models (LLMs) such as GPT [1]. In the subsequent stage, we employ a set of predefined prompt via LLMs with a fixed token length for label consolidation, to craft the final label. This process ensuring a coherent and unified representation of the hierarchical categories. Following the methodology of Sun et al. [38], we then formulate prompt pairs, which are formalized as follows:

$$\begin{aligned} \text{Prompt}^{+} &= \left[V_1^{+}, V_2^{+}, \ldots, V_{N^{+}}^{+}, \text{CLS}\right], \\ \text{Prompt}^{-} &= \left[V_1^{-}, V_2^{-}, \ldots, V_{N^{-}}^{-}, \text{CLS}\right], \end{aligned} \tag{1}$$

where $V_i^{\pm}$ represents the learnable word embedding vectors, and CLS denotes the class token, which is the generated of the Hierarchical Class Integration (HCI) process. We concatenate the CLS with learnable prompts and feed them into text encoder to obtain label embedding. Here, N^{+} and N^{-} correspond to the count of word tokens in the positive and negative prompts, respectively.

Sight-semantic Image Encoding: Recognizing the semantic divergence between visual and textual domains, we harness the extensive knowledge embedded in CLIP. However, a substantial portion of this knowledge is oriented towards sight representations, which diverges from our semantic-centric tasks. To harness this rich pre-trained knowledge while aligning it with our goals, we have augmented the CLIP architecture with a dual image encoder framework. Each encoder is tailored for specific visual interpretations and shares the same architecture and initialization weights, as illustrated in Fig. 2 (b).

The **Sight Encoder** (SiE) remains frozen during training, specifically designed to leverage the extensive sight knowledge within CLIP. It is adept at capturing tangible visual features. In contrast, the **Semantic Encoder** (SeE) is designed to be learnable. In a ResNet configuration, the initial layers, particularly the first three, are kept frozen, as they typically capture fundamental visual

elements like texture and color, which are crucial for subjective tasks. The deeper layers, more focused on high-level semantics, are configured to be adaptable for fine-tuning. This flexibility ensures that the SeE aligns with the subtleties of semantic nuances. Together, the SiE and SeE form a cohesive strategy in our IntCLIP framework. While a fixed structure offers stability, it lacks adaptability; conversely, a fully adaptable structure risks losing CLIP's sight knowledge. Our dual encoder approach deftly navigates these challenges, offering a comprehensive understanding of both tangible details and essential semantic nuances.

3.4 Sight-assisted Aggregation.

Direct concatenation or summation of sight and semantic feature maps offers a rudimentary method for harnessing the richness of two information types in intention classification. However, such a straightforward approach falls short in adequately prioritizing the hierarchy of features, particularly in terms of using sight features to enhance semantic ones. To overcome this limitation, we introduce Sight-assisted Aggregation (SA), a technique that adaptively integrates the context of sight features into the semantic feature maps. This technique underscores the supportive role of sight features in enhancing semantic feature maps, as illustrated in Fig. 4.

Semantic Query Attention: Given a semantic feature map $M_{sem} \in \mathbb{R}^{C\times H\times W}$, it undergoes convolution to yield two distinct subfeature maps:

$$\begin{aligned} M_q &= Q_{conv}(M_{sem}), \\ M_k &= K_{conv}(M_{sem}), \end{aligned} \tag{2}$$

where M_q and M_k act as query and key feature maps, respectively. Both are then reshaped to $\mathbb{R}^{C\times N}$, with $N = H\times W$ denoting pixel count. Matrix multiplication between the transposed M_q and M_k follows, with a softmax layer applied to produce the semantic attention map $M_s \in \mathbb{R}^{N\times N}$:

$$M_s(i,j) = \frac{\exp({M_q}'(i)\cdot {M_k}'(j))}{\sum_{i=1}^{N}\exp({M_q}'(i)\cdot {M_k}'(j))}, \tag{3}$$

where ${M_q}'$ is the reshaped M_q and ${M_k}'$ is the transposed, reshaped M_k. Positions with similar features exhibit stronger correlations in this map.

Sight-assisted Modeling: The sight feature map M_{sig} similarly undergoes value convolution, resulting in $M_v \in \mathbb{R}^{C\times H\times W}$. After reshaping and multiplication with the transpose of M_s, the output is reshaped back and the final intention feature map M_i is obtained by scaling with α and summing with M_{sig}:

$$M_i(j) = \alpha\sum_{i=1}^{N}(M_s(j,i)\cdot {M_v}'(i)) + M_{sig}(j), \tag{4}$$

where ${M_v}'$ is the reshaped M_v, and α, initially set to zero, is learned to assign appropriate weight during training.

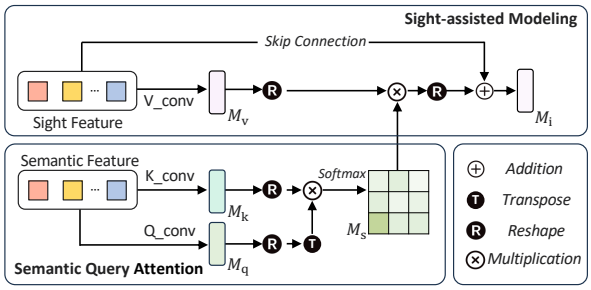

Fig. 4: Semantic Query Attention and Sight-assisted Modeling work in tandem to extract and integrate semantic query information with sight information.

From a high-dimensional perspective, the SA mechanism assigns weights to the semantic feature map, integrating sight features–particularly object-related information–into the semantic map to produce the final result. This mechanism yields M_i, a feature map in which each position represents a contextually-aware weighted combination of features. This selective aggregation process thereby enhances the semantic coherence across the entire feature map.

Optimization. After deriving the final image representation M_i and the corresponding positive/negative text representations $M_{p/n}$ from Eq. (1), we calculate the binary classification output p as:

$$p = \frac{\exp\left(\langle M_{int}, M_p \rangle / \tau\right)}{\exp\left(\langle M_i, M_p \rangle / \tau\right) + \exp\left(\langle M_{int}, M_n \rangle / \tau\right)}, \tag{5}$$

where $\langle \cdot, \cdot \rangle$ denotes the dot product, and τ is the temperature parameter that scales the logits.

Consistent with the methodology of Sun *et al.* [38], we utilize Asymmetric Loss (ASL) [30] to facilitate the optimization process in multi-label intention understanding tasks. The loss calculations for positive and negative (image, label) pairs, $\mathcal{L}_+$ and $\mathcal{L}_-$, denoted as:

$$\begin{aligned} \mathcal{L}_+ &= (1-p)^{\gamma_+} \log(p), \\ \mathcal{L}_- &= (p_c)^{\gamma_-} \log(1-p_c), \end{aligned} \tag{6}$$

with $p_c = \max(p - c, 0)$ representing the confidence-adjusted probability for negative examples, influenced by a margin c. The hyperparameters γ_- and γ_+ are adjusted to satisfy $\gamma_- \geq \gamma_+$, which allows Asymmetric Loss to effectively down-weight and apply a hard threshold to easy negative samples. The learnable prompts undergo iteratively refined through the back-propagation of Asymmetric Loss within the confines of the non-trainable text encoder.

4 Experiment

4.1 Datasets and Experiment Setups

Datasets. We evaluate the efficacy of IntCLIP in Multi-Intent Understanding (MIU) using the Intentonomy dataset, comprising 12,740 training, 498 validation, and 1,217 test images from Unsplash, categorized into 28 intent classes. IntCLIP addresses the discrepancy between subjective tasks and objective pre-training data. We further assess its performance on Image Emotion Recognition benchmarks EmotionROI_6 and FI_8 [27, 52].

Evaluation Metrics. For intention understanding, we employ Macro F1, Micro F1, and Samples F1 scores. Micro F1 calculates the unweighted mean of metrics for each label, Macro F1 aggregates metrics globally, and Samples F1 averages instance-level metrics. We assess model quality using the mean of these three metrics. For Image Emotion Recognition, we utilize classification accuracy to align with existing literature.

Implementation Details. In all baselines, we employ ResNet-101 [10] as the visual encoder and maintain an input resolution of 224×224 pixels, ensuring a fair comparison with existing CNN-based methods [38, 50]. The Transformer architecture [29, 39] from CLIP [28] serves as our text encoder. Both the visual and textual encoders, initialized from the CLIP model, are configured such that the Sight Encoding branch and the text encoder remain fixed during training. For each class or label, two independent context vectors are learned, each possessing 16 context tokens ($N = 16$) following [57]. Training employs the SGD optimizer initialized at a learning rate of 0.002, which is adaptively decayed according to the cosine annealing schedule. The context vectors are trained for 50 epochs, using batch sizes of 32 and 16 for intention and emotion tasks, respectively. The ASL parameters, $\gamma^+ = 1$, $\gamma^- = 2$, and $c = 0.05$, are selected based on prior multi-label classification research [38].

Baselines. To underscore the effectiveness of IntCLIP, we benchmark it against the following methods:

1. PIP-Net [41], CPAD [50], HLEG [33], and the original method within Intentonomy [17] are tailored for the MIU task. Their performances in the MIU context are reported as per the strategies described in [33, 41].
2. MultiGuard [16], SST [3] Query2Label [24] and DualCoOp [38] are crafted for multi-label tasks. Specifically, DualCoOp leverages large-scale pre-trained knowledge in multi-label recognition, employing a dual context prompting strategy to enhance classification accuracy.
3. In the field of image-based emotion recognition task, we include ECWA [5], WSEIP [53], CGLF-Net [25] and MSRCA [54] significant influence in the field. With IntCLIP, we preserve consistent knowledge foundations across both intention and emotion recognition, offering a comprehensive comparative analysis.

Table 1: Performance comparison of various multi-label classification and intention understanding methods on the Intentonomy dataset, subdivided into Testing (2K images) and Validation (2K images) sets. #P indicates the number of learnable parameters. Our IntCLIP models, with ResNet101 (R101) and Vision Transformer (ViT-B/32) backbones, demonstrate superior performance. Gray indicates use of a more powerful feature extractor (ViT).

Methods	#P	Intentonomy Testing (2K)				Intentonomy Validation (2K)			
		MacroF1	MicroF1	SamplesF1	mAP	MacroF1	MicroF1	SamplesF1	mAP
Multi-label classification methods									
MultiGuard (Jia et al., 2022) [16]	53.6M	26.87	38.37	40.06	36.08	27.74	41.03	41.58	37.55
SST (Chen et al., 2022) [3]	33.5M	27.33	42.21	42.62	35.93	28.78	41.03	41.75	32.12
Query2Label (Liu et al., 2021) [24]	82.4M	32.12	44.64	45.15	34.33	35.35	46.34	47.40	36.86
DualCoOp (Sun et al., 2022) [38]	14.9M	31.06	44.42	40.40	39.33	32.67	46.29	42.97	40.93
Intention Understanding methods									
Intentonomy (Jia et al., 2021) [17]	46.5M	23.98	31.28	31.39	-	25.07	32.94	33.61	-
CPAD (Ye et al., 2023) [50]	46.7M	27.39	40.98	41.12	-	27.37	41.77	42.68	-
HLEG (Shi et al., 2023) [33]	89.4M	32.77	44.69	45.54	35.93	35.35	46.34	47.40	36.86
PIP-Net (Wang et al., 2023) [41]	46.3M	32.57	45.94	47.05	-	30.71	44.84	45.48	-
LabCR (Shi et al., 2024) [32]	56.2M	34.63	48.51	48.05	37.13	37.10	49.04	49.71	38.86
Our IntCLIP (R101)	15.7M	38.40	50.54	49.31	42.66	40.09	50.68	48.99	42.84
Our IntCLIP (ViT-B/32)	22.6M	40.15	53.54	51.01	45.32	41.55	52.61	49.75	44.95

4.2 Multi-Label Intention Understanding

To evaluate the effectiveness of our proposed IntCLIP model, we undertook a comprehensive comparative study on the Intentonomy dataset. This dataset includes both a Testing and a Validation set. The comparative performance results, as shown in Tab. 1, indicate how various state-of-the-art models fare against our IntCLIP. We categorized the methods into two groups: (i) general multi-label classification methods, and (ii) methods specialized in multi-label intention understanding.

While general multi-label classification models like DualCoOp [38] have shown promise, specialized methods for intention understanding, such as HLEG [33] and PIP-Net [41], have demonstrated superior performance. Our proposed IntCLIP model surpasses the state-of-the-art, achieving a Micro F1 score of 50.54% on the test set, an 11.2% improvement over previous best-performing models. This advancement can be attributed to the synergistic integration of Hierarchical Class Integration, Sight-semantic Image Encoding, and CLIP's foundational cross-modal alignment knowledge. These components collectively enhance semantic understanding of intentions, capture both objective and subjective nuances, and ensure robust general cross-modal alignment.

It's noteworthy that IntCLIP achieves rapid adaptation to intention tasks with only 15.7 million learnable parameters for the ResNet101 (R101) backbone. This is notably lower than that of competing methods, highlighting IntCLIP's efficiency[3]. Despite having 14.2 million additional parameters during inference due to the extra semantic layers and aggregation module, this increase is deemed worthwhile for boosting in understanding capacity. IntCLIP's leaner architecture

[3] For baselines without public released implementation, we estimate the major portion of the learnable parameters based on descriptions in their respective papers.

Table 2: Performance analysis when excluding individual components: Hierarchical Class Integration (HCI), Sight-semantic Aggregation (SA), and Sight-semantic Image Encoding (SIE). Their combined use boosts the model's overall performance.

w/o Component	Intentonomy Testing (2K)			
	SamplesF1	MacroF1	MicroF1	mAP
w/o HCI	44.74	36.18	47.05	39.81
w/o SA	47.25	35.15	47.46	39.42
w/o SIE,SA	44.71	35.52	45.91	39.02
Full IntCLIP	**49.31**	**38.40**	**50.54**	**42.66**

underscores its ability to deliver enhanced performance while ensuring computational efficiency.

4.3 Ablation Studies

Effectiveness of Proposed Components. Our framework integrates three key components: Hierarchical Class Integration (HCI), Sight-assisted Aggregation (SA), and Sight-semantic Image Encoding (SIE). While each component contributes to performance enhancement, as evidenced in Tab. 2, it is their combined synergy that culminates in best performance. The comprehensive integration of HCI, SA, and SIE effectively maximizes alignment knowledge, endowing the model with greater versatility and precision.

Optimizing Hierarchical Class Integration with LLMs. We investigated various hierarchical label integration strategies for multi-label intention understanding on the Intentonomy Testing dataset (Tab. 3). The baseline method using only 'Fine-grained' labels achieved a Macro F1 of 36.04% and Micro F1 of 46.46%. While the 'Consolidation' technique marginally improved these metrics, it failed to fully leverage the hierarchical semantic structure. The Multi Layer Transformer, inspired by HLEG [33], showed only slight improvements, likely due to information and parameter redundancies.

Our proposed method significantly enhanced performance across all metrics with varying content lengths. Optimal performance was achieved with 11-15 words, yielding a Macro F1 of 38.40%, Micro F1 of 50.54%, and mAP of 42.66%. This demonstrates the efficacy of Large Language Models (LLMs) in generating coherent and context-rich hierarchical labels. However, increasing content length to 16-20 words led to mixed results, suggesting that overly verbose labels may introduce superfluous information.

Our Hierarchical Class Integration (HCI) method effectively harnesses the semantic richness of hierarchical labels while avoiding redundancy issues, establishing a new benchmark on the Intentonomy dataset and highlighting the potential of LLMs in multi-label intention understanding tasks.

Sight-semantic Image Encoding Design. Sight-semantic Image Encoding (SIE) plays a crucial role in aligning sight knowledge with the demands of semantic tasks. An important observation from Tab. 4 is that fully training the

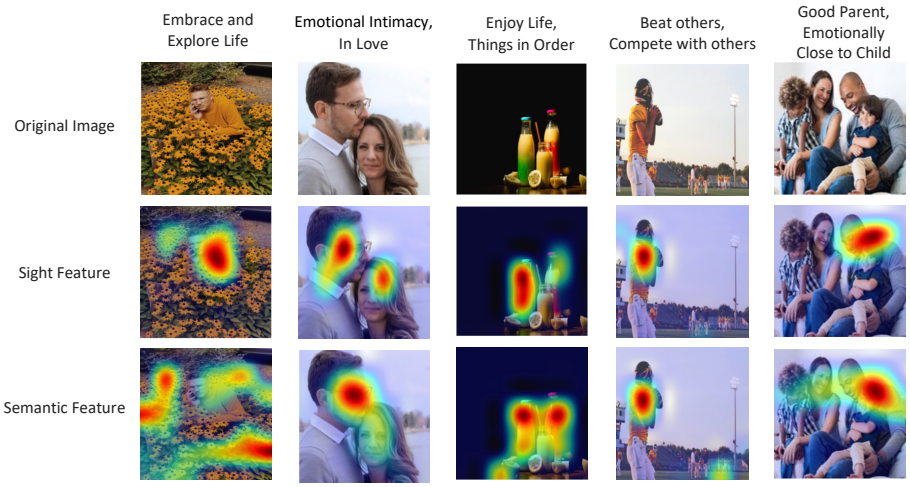

Fig. 5: Encoder Focus Visualization in Sight-semantic Image Encoding. Original images (top) with their intent labels and corresponding heatmaps from the sight (middle) and semantic (bottom) encoders. The semantic encoder's heatmaps show its alignment with intent-relevant areas, demonstrating its enhanced capability.

visual feature extractor, whether in a single or dual-branch approach, often leads to conflicts between sight knowledge and semantic alignment, resulting in diminished performance. In the single-branch setup, confining training solely to semantic layers noticeably reduces the utilization of pre-trained knowledge. In contrast, our dual-branch approach, which restricts training to the semantic layers, successfully navigates this challenge. This finding highlights our design's effectiveness, validating the dual-branch configuration as a more suitable solution for balancing pre-trained knowledge with task-specific requirements.

To further substantiate the efficacy of the Sight-semantic Encoding approach, we utilized GradCAM visualization as illustrated in Fig. 5 [31]. The capability of the semantic feature to identify relevant aspects within visual data, thereby improving understanding, is clearly demonstrated across the first, second, third, and fifth columns. Additionally, attention is drawn to the network's emphasis on other participants in the playing area in the fourth column, which aids in the deduction of 'Beat others' within the label. The comparative analysis underscores that, whereas the sight encoder targets visually prominent regions, the semantic encoder zeroes in on zones that bear a closer relation to intention-specific attributes. This differentiation is crucial for accurate intention recognition, highlighting the semantic encoder's proficiency to link visual features with underlying intentions.

Aggregation Function Analysis. The choice of aggregation function significantly influences the model's capacity to meld information from different sources. Simple element-wise addition fails to sufficiently refine sight cue components, offering inadequate guidance for semantic features. The Concat + MLP approach,

Table 3: Comparison of Hierarchical Class Integration (HCI) by Large Language Models (LLMs) with varying context lengths against baseline methods on the Testing set. The results demonstrate that HCI, especially with context lengths of 11-15 and 16-20, outperforms fine-grained label, consolidated label and multi-layer methods, indicating the effectiveness of sentence-level prompt in improving intention understanding.

Methods	Context Length	Intentonomy Testing Set (2K)			
		MacroF1	MicroF1	SamplesF1	mAP
Fine-Grained	-	36.04	46.46	44.48	38.44
Consolidation	-	36.18	47.05	44.74	39.81
Multi Layer [33]	-	36.52	46.96	44.98	38.81
HCI by LLM	3-5	37.02	47.98	45.63	40.35
	6-9	38.68	49.43	46.67	42.34
	11-15	**38.40**	50.54	**49.31**	**42.66**
	16-20	37.76	**50.80**	46.05	42.03

Table 4: Comparison of various encoder designs. This showcases the conflicts arising from training all layers and the superiority of Sight-semantic Image Encoding.

Various Visual Encoder Designs		Intentonomy Testing (2K)			
Brunch	Training Layer	MacroF1	MicroF1	SamplesF1	mAP
Single	✗	34.53	43.84	43.43	38.63
Single	All	9.85	23.18	19.95	12.42
Single	Semantic	35.76	46.83	44.25	39.88
Dual	✗	36.85	47.73	44.99	40.35
Dual	All	6.28	22.64	18.89	14.02
Dual	Semantic	**38.40**	**50.54**	**49.31**	**42.66**

Table 5: Analysis of different aggregation methods. Our Sight-assisted Aggregation (SA) offers a more judicious selection of information, leading to enhanced aggregation.

Aggragation Methods	Intentonomy Testing Set (2K)			
	MacroF1	MicroF1	SamplesF1	mAP
Element-wise Addition	35.15	47.46	47.25	39.42
Concat + MLP	35.30	46.94	46.85	40.51
Ours	**38.40**	**50.54**	**49.31**	**42.66**

while more intricate, risks missing key information during the MLP phase. Our SA approach excels by selectively integrating sight features, achieving a more comprehensive aggregation, as demonstrated in Tab. 5.

Generalization to Image Emotion Recognition. A critical aspect of our approach involves recognizing and addressing the disparities between sight and semantic image-text pairs in subjective tasks. Image-based Emotion Recognition (IER) falls squarely within this scope. To validate the generalization of our method, we conducted evaluations on the EmotionROI_6 and FI_8 datasets. The results, as illustrated in Tab. 6 (a), not only demonstrate our method's ability to bridge the disparity but also establish new benchmarks in IER. These

Table 6: Generalization analysis of IntCLIP and HCI. (a) Performance on Image Emotion Recognition (IER) task. (b) Effectiveness of HCI. HCI–FG indicates that only with fine-grained labels.

(a) Generalization to IER.

Methods	Emo6	FI8
ECWA [5]	59.09	70.87
WSEIP [53]	60.41	75.91
MSRCA [54]	55.60	72.60
CGLF-Net [25]	65.01	75.61
IntCLIP (R101)	**71.05**	**78.93**

(b) Generalization of HCI.

Methods	Macro F1	Micro F1	Samples F1
HCI	41.55	52.61	49.75
HCI-FG	41.33	52.14	50.14
[33]	35.35	46.34	47.40
[33]+HCI	36.47	48.20	48.64

outcomes highlight the robustness of IntCLIP in subjective task domains. We believe the insights on combination of sight and semantic information would be beneficial to other high-level semantic understanding tasks.

Generalization of HCI. The results in Tab. 6 (b) highlight two key aspects of HCI. HCI-FG represents an exploration of using only fine-grained labels to generate integrated labels, demonstrating the method's adaptability. The comparable performance of HCI-FG to standard HCI showcases the broad applicability of HCI. Additionally, the improvement when HCI is combined with other approaches, such as HLEG, underscores its effectiveness as a general technique.

5 Conclusion

We addresse the complexities of Multi-label Intention Understanding (MIU) in social media imagery, where intentions often transcend visual elements. We introduce IntCLIP, a model designed to overcome the challenge of limited annotated data by synergizing sight knowledge with semantic intention cues. The key features in IntCLIP include Sight-semantic Image Encoding, Hierarchical Class Integration, and Sight-assisted Aggregation, enhancing its capacity for deep understanding in MIU tasks. Empirical evaluations demonstrate the superior performance of IntCLIP over established techniques in MIU benchmarks and specialized tasks such as Image Emotion Recognition. By leveraging large-scale multimodal pre-training, IntCLIP significantly advances the field, enabling more nuanced interpretations of intentions in social media imagery and contributing to the broader understanding of digital communication.

Acknowledgment This work is supported by the National Natural Science Foundation of China under Grant (62361166629, 62176188). The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NIPS (2020)
2. Chen, S., Ye, M., Du, B.: Rotation invariant transformer for recognizing object in uavs. In: ACM MM (2022)
3. Chen, T., Pu, T., Wu, H., Xie, Y., Lin, L.: Structured semantic transfer for multi-label recognition with partial labels. In: AAAI (2022)
4. Chenyue Li, S.C., Ye, M.: Adaptive high-frequency transformer for diverse wildlife re-identification. In: ECCV (2024)
5. Deng, S., Wu, L., Shi, G., Zhang, H., Hu, W., Dong, R.: Emotion class-wise aware loss for image emotion classification. In: Artificial Intelligence: First CAAI International Conference (2021)
6. Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., Hon, H.W.: Unified language model pre-training for natural language understanding and generation. NIPS (2019)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2020)
8. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International Journal of Computer Vision (2023)
9. Ghaisani, A.P., Handayani, P.W., Munajat, Q.: Users' motivation in sharing information on social media. Procedia Computer Science (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
11. Huang, T., Chu, J., Wei, F.: Unsupervised prompt learning for vision-language models. arXiv:2204.03649 (2022)
12. Huang, W., Ye, M., Shi, Z., Du, B.: Generalizable heterogeneous federated cross-correlation and instance similarity learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
13. Huang, W., Ye, M., Shi, Z., Li, H., Du, B.: Rethinking federated learning with domain shift: A prototype view. In: CVPR (2023)
14. Huang, W., Ye, M., Shi, Z., Wan, G., Li, H., Du, B., Yang, Q.: A federated learning for generalization, robustness, fairness: A survey and benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
15. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML (2021)
16. Jia, J., Qu, W., Gong, N.: Multiguard: Provably robust multi-label classification against adversarial examples. NIPS (2022)
17. Jia, M., Wu, Z., Reiter, A., Cardie, C., Belongie, S., Lim, S.N.: Intentonomy: a dataset and study towards human intent understanding. In: CVPR (2021)
18. Joo, J., Li, W., Steen, F.F., Zhu, S.C.: Visual persuasion: Inferring communicative intents of images. In: CVPR (2014)
19. Joo, J., Steen, F.F., Zhu, S.C.: Automated facial trait judgment and election outcome prediction: Social dimensions of face. In: ICCV (2015)
20. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence (2013)

21. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: EMNLP (2021)
22. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: IJCNLP (2021)
23. Liu, B., Adeli, E., Cao, Z., Lee, K.H., Shenoi, A., Gaidon, A., Niebles, J.C.: Spatiotemporal relationship reasoning for pedestrian intent prediction. IEEE Robotics and Automation Letters (2020)
24. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Query2label: A simple transformer way to multi-label classification. arXiv:2107.10834 (2021)
25. Luo, Y., Zhong, X., Zeng, M., Xie, J., Wang, S., Liu, G.: Cglf-net: Image emotion recognition network by combining global self-attention features and local multiscale features. IEEE Transactions on Multimedia (2023)
26. Newall, N., Smith, B.G., Burton, O., Chari, A., Kolias, A.G., Hutchinson, P.J., Alamri, A., Uff, C., Adegboyega, G., Ali, M., et al.: Improving neurosurgery education using social media case-based discussions: a pilot study. World Neurosurgery: X (2021)
27. Peng, K.C., Sadovnik, A., Gallagher, A., Chen, T.: Where do emotions come from? predicting the emotion stimuli map. In: ICIP (2016)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
29. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog (2019)
30. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., Zelnik-Manor, L.: Asymmetric loss for multi-label classification. In: ICCV (2021)
31. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-cam: Why did you say that? arXiv:1611.07450 (2016)
32. Shi, Q., Ye, M., Huang, W., Ruan, W., Du, B.: Label-aware calibration and relation-preserving in visual intention understanding. IEEE Transactions on Image Processing (2024)
33. Shi, Q., Ye, M., Zhang, Z., Du, B.: Learnable hierarchical label embedding and grouping for visual intention understanding. IEEE Transactions on Affective Computing (2023)
34. Shi, W., Ye, M.: Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning. In: ICCV (2023)
35. Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In: EMNLP (2020)
36. Stavros, C., Meng, M.D., Westberg, K., Farrelly, F.: Understanding fan motivation for interacting on social media. Sport management review (2014)
37. Suarez-Lledo, V., Alvarez-Galvez, J.: Prevalence of health misinformation on social media: systematic review. Journal of Medical Internet Research (2021)
38. Sun, X., Hu, P., Saenko, K.: Dualcoop: Fast adaptation to multi-label recognition with limited annotations. NIPS (2022)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. NIPS (2017)
40. Vondrick, C., Oktay, D., Pirsiavash, H., Torralba, A.: Predicting motivations of actions by leveraging text. In: CVPR (2016)
41. Wang, B., Yang, K., Zhao, Y., Long, T., Li, X.: Prototype-based intent perception. IEEE Transactions on Multimedia (2023)

42. Wang, S., Liu, F., Yu, M.: Persuasion knowledge and consumers' attitudes towards benefit-based advertising. In: Australia and New Zealand Marketing Academy Conference Proceedings (2012)
43. Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., Yu, Z.: Persuasion for good: Towards a personalized persuasive dialogue system for social good. In: ACL (2019)
44. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: CVPR (2022)
45. Wu, J., Zheng, R., Jiang, J., Tian, Z., Chen, W., Wang, Z., Yu, F.R., Leung, V.C.M.: A lightweight small object detection method based on multilayer coordination federated intelligence for coal mine iovt. IEEE Internet of Things Journal (2024)
46. Yang, Q., Ye, M., Cai, Z., Su, K., Du, B.: Composed image retrieval via cross relation network with hierarchical aggregation transformer. IEEE Transactions on Image Processing (2023)
47. Yang, Q., Ye, M., Du, B.: Emollm: Multimodal emotional understanding meets large language models (2024)
48. Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: Cpt: Colorful prompt tuning for pre-trained vision-language models. arXiv:2109.11797 (2021)
49. Ye, M., Chen, S., Li, C., Zheng, W.S., Crandall, D., Du, B.: Transformer for object re-identification: A survey. arXiv:2401.06960 (2024)
50. Ye, M., Shi, Q., Su, K., Du, B.: Cross-modality pyramid alignment for visual intention understanding. IEEE Transactions on Image Processing (2023)
51. Ye, M., Wu, Z., Chen, C., Du, B.: Channel augmentation for visible-infrared re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
52. You, Q., Luo, J., Jin, H., Yang, J.: Building a large scale dataset for image emotion recognition: The fine print and the benchmark. In: AAAI (2016)
53. Zhang, H., Xu, M.: Weakly supervised emotion intensity prediction for recognition of emotions in images. IEEE Transactions on Multimedia (2020)
54. Zhang, J., Liu, X., Chen, M., Ye, Q., Wang, Z.: Image sentiment classification via multi-level sentiment region correlation analysis. Neurocomputing (2022)
55. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv:2111.03930 (2021)
56. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: CVPR (2022)
57. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision (2022)
58. Zyner, A., Worrall, S., Ward, J., Nebot, E.: Long short term memory for driver intent prediction. In: 2017 IEEE Intelligent Vehicles Symposium (2017)