Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SymCLIP: Symmetric Prompting with Adaptive Semantic Labeling for Multi-Intent Recognition

Shuang Wu
Henan University of Technology
Honglin Ma
mhl@haut.edu.cn

Henan University of Technology

Research Article

Keywords: Multi-label, Intention Understanding, CLIP Adaptation, Vision, Language

Posted Date: September 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10542927/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SymCLIP: Symmetric Prompting with Adaptive Semantic Labeling for Multi-Intent Recognition

Shuang Wu[1] and Honglin Ma[2*]

[1]School of Information Science and Engineering, Henan University of Technology, Lianhua Street, Zhengzhou, 450000, Henan Province, China.
[2*]School of Information Science and Engineering, Henan University of Technology, Lianhua Street, Zhengzhou, 450000, Henan Province, China.

*Corresponding author(s). E-mail(s): mhl@haut.edu.com; wshuang783@gmail.com; mhl@haut.edu.com;

Abstract

Understanding intent in social media images is challenging in Multi-label Intention Understanding (MIU) due to the ambiguity and abstractness of semantic intentions. Although vision–language models like CLIP provide strong foundations, existing adaptation methods often fail to establish an explicit and synergistic connection between visual features and textual intent labels during the learning process, treating them in isolation. Moreover, the use of fixed intent categories lacks adaptability to dynamic and context-dependent expression scenarios, limiting the model's interpretative flexibility. In this paper, we propose SymCLIP, a framework that enhances intent understanding through Symmetric Vision–Language Prompting (SVLP) and Dynamic Label Integration (DLI). SVLP couples the two modalities by generating visual prompts from language prompts, enabling semantically guided visual feature extraction. DLI replaces fixed class labels with trainable embeddings, allowing adaptive and fine-grained intent representation. Experiments on the Intentonomy dataset show that SymCLIP achieves a new state-of-the-art MicroF1 of 55.38%, outperforming previous methods by 9.6%. Additional tests on image emotion recognition further confirm its strong generalization across subjective visual understanding tasks.

Keywords: Multi-label, Intention Understanding, CLIP Adaptation, Vision–Language

1 Introduction

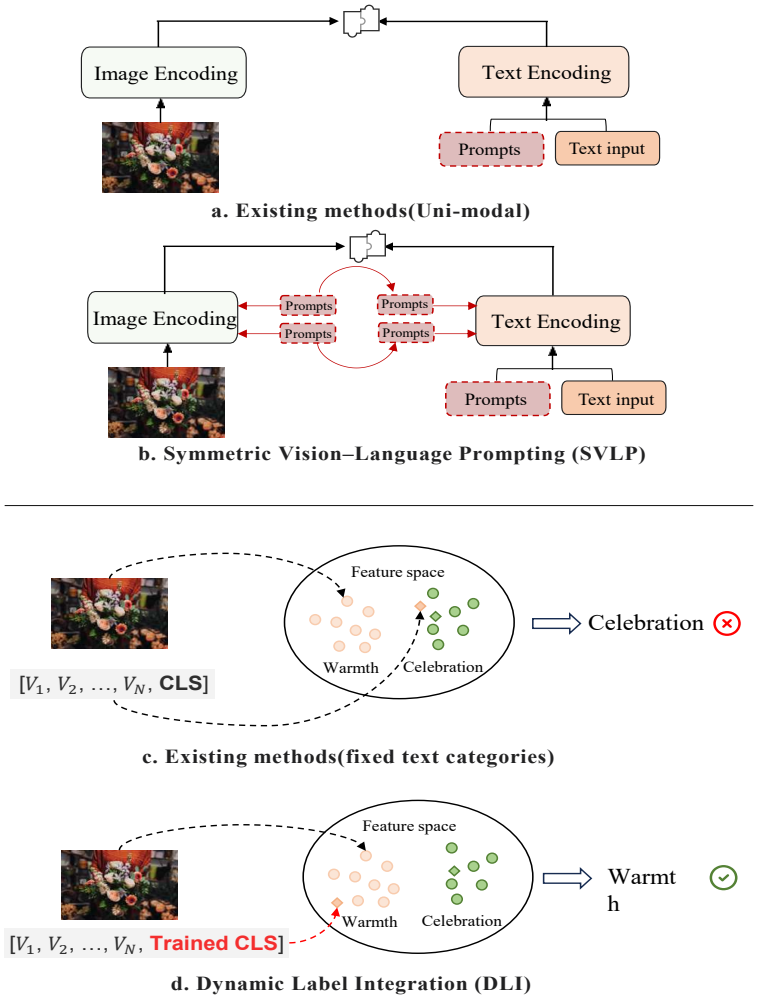


Fig. 1 Limitations of existing MIU methods and proposed improvements. (a) Conventional approaches use unimodal prompt tuning on a single CLIP branch, limiting cross-modal alignment. (b) SVLP introduces branch-aware hierarchical prompts for joint adaptation of visual and language branches. (c) Fixed text categories in prior methods restrict flexibility in capturing diverse intents. (d) DLI automatically optimizes text embeddings via gradient-based search for adaptive intent representation.

In today's digital landscape, images on social media go beyond mere visual representation, often carrying implicit intentions. Whether used to influence, commemorate, or convey information, these images play a crucial role in areas such as monitoring mental well-being [1] and addressing misinformation [2], emphasizing the need

for deeper recognition of intent. Hence, Multi-label Intention Understanding (MIU) represents a fundamental yet complex challenge in the domain of image analysis. In recent years, with the continuous advancement of deep learning technologies [3–5] and their widespread applications in real-world scenarios, research on understanding human intentions has garnered increasing attention. Meanwhile, pre-trained vision-language (V-L) models, such as CLIP [6], have demonstrated remarkable generalization capabilities in standard image classification tasks. However, their learned knowledge is fundamentally vision-oriented, prioritizing low-level visual cues such as object contours, color information, and spatial layout. This emphasis allows for effective recognition of observable image content, but offers limited capacity to capture the deeper semantic factors essential for MIU tasks. In particular, elements such as contextual relevance, cultural interpretation and emotional implication may be implicitly derived from visual input, but they extend beyond what can be captured through superficial appearance alone. Consequently, existing models in the style of CLIP still exhibit a considerable gap from the level of semantic understanding required for MIU tasks.

To bridge this gap, Yang et al. proposed IntCLIP [7], which for the first time applied CLIP's visual knowledge to the MIU domain, achieving precise capture of semantic intent. However, existing methods do not establish the **explicited connection in the prompts of images and class labels** during the learning process, but learn separately (Fig. 1:a). This weakens the association between intent categories and corresponding images during training. In addition, fixed text categories **cannot adapt to flexible intent understanding scenarios**, which affects the understanding ability of the model(Fig. 1:c).

To overcome these challenges, we propose Symmetric Vision-languege Prompting CLIP (SymCLIP). Specifically, in order to establish an explicit connection between image and text prompts during training, we propose the Symmetric Vision-Language Prompting (SVLP) method(Fig. 1:b), which integrates visual and textual prompts through a symmetric way to achieve the synergy between the vision and language branches in the intention understanding task. Moreover, to enhance the model's adaptability to flexible intent understanding scenarios and improve its intent recognition capabilities in MIU tasks, we introduce the Dynamic Label Integration (DLI) technique(Fig. 1:d), which optimizes the $\langle$CLASS$\rangle$ token embeddings through combining the label text with trainable vectors. We conducted experiments on the Intentonomy dataset, which encompasses 28 intent categories, to validate the effectiveness of our approach in MIU tasks. Through the seamless integration of SVLP and DLI, our method not only effectively manages computational overhead but also demonstrates superior efficiency in both training and inference phases, along with advantages in continual learning. Compared to IntCLIP [7], our approach achieves a substantial improvement in mean average precision across categories. SymCLIP establishes new state-of-the-art performance on the Intentonomy dataset, achieving a Micro-F1 score of 55.38%, a Macro-F1 of 45.06%, and a mean Average Precision (mAP) of 49.26% on the testing set, surpassing the previous best IntCLIP model by 9.6%, 4.9%, and 3.9%, respectively.

To summary, our contribution are as follw:

- We propose SymCLIP, a new framework for Multi-label Intention Understanding that explicitly addresses the challenge of aligning visual features with abstract semantic intents by establishing a tighter coupling between image and text modalities.
- We introduce Symmetric Vision-Language Prompting (SVLP), a novel mechanism that creates a synergistic link between the vision and language branches. Unlike prior work that optimizes prompts independently , SVLP dynamically generates visual prompts from textual ones, ensuring the visual encoder is explicitly guided by semantic intent during training.
- We present the Dynamic Label Integration (DLI) technique, which replaces fixed textual categories with trainable class token embeddings. This allows the model to learn more nuanced and flexible representations of intent categories directly from data, enhancing its adaptability and recognition capabilities in complex scenarios.
- We conduct extensive experiments on the Intentonomy benchmark, where SymCLIP achieves new state-of-the-art performance across multiple metrics, outperforming prior methods by a significant margin. We also demonstrate the model's strong generalization to other subjective perception tasks, such as image emotion recognition, validating the robustness of our approach

The rest of this paper is organized as follows: Section 2 introduces the background, including fundamental concepts of vision-language representation learning and prompt learning. Section 3 reviews the related work. Section 4 presents the proposed SymCLIP framework, including the Symmetric Vision-Language Prompting (SVLP) mechanism and the Dynamic Label Integration (DLI) strategy. Section 5 provides extensive experiments, including comparative evaluations, ablation studies, and cross-task generalization analysis. Finally, Section 6 concludes the paper and discusses future research directions.

2 Background

2.1 Foundations of Multi-label Intention Understanding

Multi-label Intention Understanding (MIU) aims to infer multiple latent semantic intentions from a single image, representing a challenging task that combines multi-label learning with high-level semantic reasoning. Unlike conventional image classification, which focuses on objective and observable categories, MIU requires the model to capture implicit semantic information such as emotional expression, behavioral motivation, and social context. These semantics are often subjective and ambiguous, significantly increasing the complexity of the task, as illustrated in Fig. 2

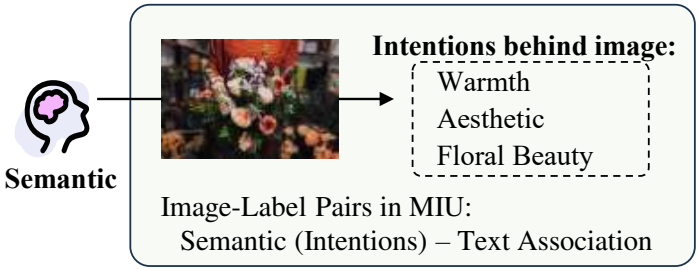


Fig. 2 Semantic Ambiguity and Annotation Complexity in MIU task. Multi-label Intention Understanding (MIU) focuses on capturing nuanced and often ambiguous semantic cues that reflect users' underlying intentions, making annotation challenging and labor-intensive.

Moreover, the labels in MIU are typically non-exclusive and exhibit strong correlations. A single image may simultaneously convey multiple intentions, such as social interaction and emotional expression, which are semantically related. Therefore, the model is required not only to distinguish individual categories but also to model the dependencies among labels for a comprehensive understanding.

From a data perspective, MIU datasets often follow a long-tailed distribution, where certain categories are underrepresented. This imbalance makes it difficult for the model to learn robust representations for rare classes. In addition, since intention annotations are based on human interpretation, discrepancies among annotators can introduce noise and inconsistency into the data set. Together, these factors make MIU a complex and challenging problem.

2.2 Vision-Language Representation Learning

Vision-Language (V-L) representation learning aims to bridge visual and textual modalities by projecting them into a shared embedding space. Typically, an image encoder $g(\cdot)$ and a text encoder $\phi(\cdot)$ are employed to extract visual and textual features, respectively. The similarity between an image $\mathcal{I}$ and a text description h_i is commonly measured using cosine similarity:

$$\text{sim}(\mathcal{I}, h_i) = \cos(g(\mathcal{I}), \phi(h_i)). \tag{1}$$

To obtain a probabilistic interpretation, a softmax function is applied over all candidate classes:

$$P(y = i \mid \mathcal{I}) = \frac{\exp(\text{sim}(\mathcal{I}, h_i)/\gamma)}{\sum_{r=1}^{K} \exp(\text{sim}(\mathcal{I}, h_r)/\gamma)}, \tag{2}$$

where γ denotes the temperature parameter that controls the sharpness of the distribution.

Recent large-scale pre-trained V-L models [8, 9] have demonstrated strong generalization capabilities by leveraging massive image-text pairs and contrastive learning objectives. By encouraging matched image-text pairs to have higher similarity than mismatched ones, these models learn a globally aligned embedding space that supports flexible cross-modal reasoning. Such a training paradigm enables efficient knowledge transfer to various downstream tasks, particularly under zero-shot and few-shot settings.

Nevertheless, the learned alignment is predominantly driven by coarse-grained semantic correspondence, which tends to emphasize visually salient and linguistically explicit concepts. As a result, the representations are often biased toward observable attributes such as object categories, spatial layouts, and scene compositions. While effective for conventional recognition tasks, this bias limits the model's capacity to encode more abstract and implicit semantics.

In scenarios involving high-level understanding, such as intention and emotion recognition, semantic cues are often subtle, context-dependent, and not directly grounded in explicit visual patterns. Consequently, the inherent gap between low-level visual features and high-level semantic concepts remains insufficiently addressed by existing V-L models. This discrepancy highlights the need for more task-adaptive mechanisms that can refine cross-modal alignment and enhance the representation of abstract semantics.

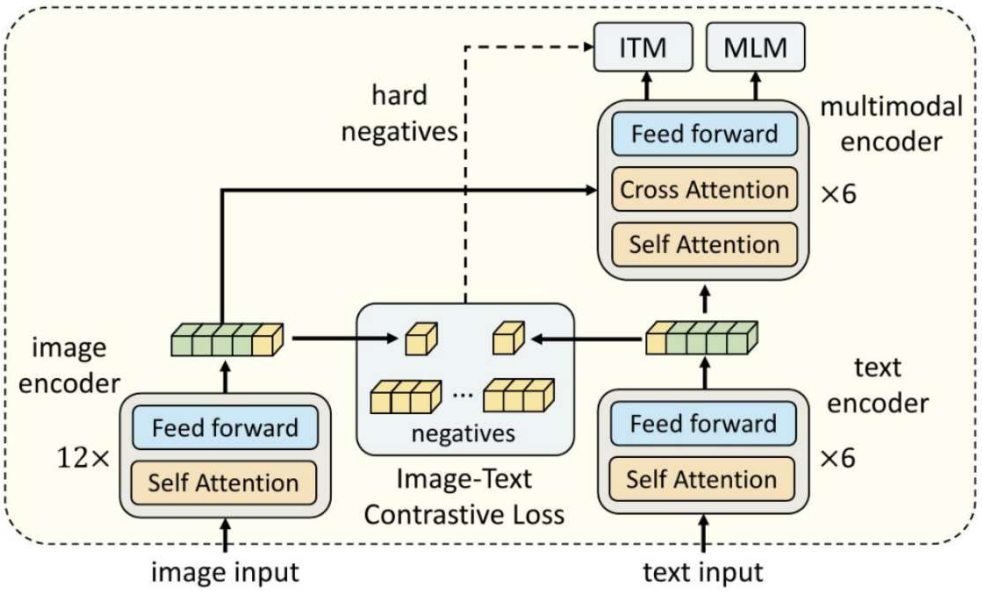


Fig. 3 Typical vision–language (V-L) model framework

2.3 Prompt Learning for Vision-Language Models

Prompt learning has emerged as an effective paradigm for adapting pre-trained V-L models to downstream tasks, offering a flexible alternative to conventional fine-tuning strategies. Instead of updating the entire set of model parameters, prompt learning introduces a small number of learnable tokens that serve as task-specific guidance signals. These tokens are optimized to reshape the input representation, enabling the model to better capture task-relevant semantics while preserving the general knowledge encoded in the pre-trained backbone. Such a mechanism allows for efficient adaptation without disrupting the intrinsic cross-modal alignment learned during large-scale pre-training.

In the textual branch, prompt learning operates by augmenting the input sequence with a set of context tokens that precede the class token. These context tokens implicitly encode task-specific semantics and guide the text encoder toward more discriminative representations. Compared to using fixed, manually designed templates, learnable prompts can dynamically adjust to the underlying data distribution, thereby reducing the reliance on human prior knowledge and improving adaptability across different tasks. Moreover, since the prompt parameters are relatively lightweight, the optimization process is more stable and less prone to overfitting, especially in scenarios with limited labeled data.

In the visual branch, prompt learning is typically realized by injecting learnable tokens into the visual input sequence or intermediate feature layers. These visual prompts act as additional carriers of task-specific information, influencing the feature extraction process by modulating attention distributions within the encoder. By interacting with image patches at different levels, visual prompts can guide the model to focus on regions that are more relevant to the downstream objective, thus enhancing the discriminative power of visual representations. This design is particularly beneficial for tasks where critical information is sparsely distributed or requires selective attention.

Despite these advantages, most existing prompt learning approaches[10–12] adopt a unimodal optimization paradigm, in which textual and visual prompts are learned independently within their respective branches. This design implicitly assumes that each modality can be optimized in isolation, without considering the mutual dependency between visual and linguistic information. However, in vision-language models, the final prediction is inherently determined by the interaction between the two modalities. Ignoring this interaction during the prompt learning process may lead to suboptimal alignment, especially when dealing with complex semantic tasks that require fine-grained correspondence between visual cues and textual concepts.

Furthermore, independent optimization may result in inconsistencies between the representations learned by different modalities. For instance, textual prompts may emphasize certain semantic attributes that are not adequately reflected in visual prompts, leading to a mismatch in the cross-modal similarity computation. This issue becomes more pronounced in tasks that involve abstract or high-level semantics, where the correspondence between modalities is less explicit and more dependent on contextual reasoning. Therefore, a more principled approach that explicitly models

the interaction between visual and textual prompts is necessary to fully exploit the potential of vision-language models.

These observations suggest that prompt learning should not be treated as a purely modality-specific adaptation mechanism, but rather as a joint optimization process that considers cross-modal dependencies. By enabling coordinated learning between visual and textual prompts, it is possible to establish a more consistent and semantically aligned representation space, thereby improving the model's ability to capture complex relationships and enhancing its performance on challenging downstream tasks.

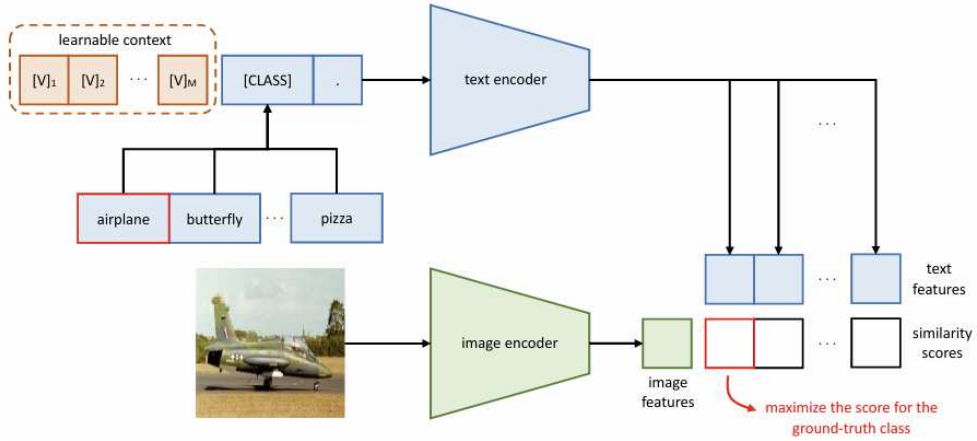


Fig. 4 Prompt Learning for Vision-Language Models

3 Related Work

3.1 Vision Language Models

Vision-Language (V-L) models have witnessed substantial advancements in aligning visual and textual representations in the past decade. Notable models such as CLIP [6], ALIGN [13], LiT [14], FILIP [15], and Florence [16] have exhibited strong generalization capabilities, achieving impressive results in diverse tasks such as few-shot and zero-shot visual recognition. These models are typically trained on large-scale image-text pairs using contrastive learning objectives, which encourage semantically aligned visual and textual representations to be close in a shared embedding space while pushing apart mismatched pairs. Such a training paradigm enables flexible cross-modal transfer and strong generalization to unseen categories. Nevertheless, efficiently leveraging these pre-trained models for downstream applications continues to pose a substantial challenge, as their representations are often biased toward generic visual semantics rather than task-specific requirements. This has prompted a surge of follow-up research, proposing tailored strategies to enhance V-L model adaptation for tasks such as few-shot image recognition [17–19], object detection [20–25], and semantic

segmentation [26–29], where specific modifications are introduced to bridge the gap between global alignment and fine-grained prediction.

To improve the model's task comprehension, prompt learning was first introduced in the NLP [30–32] domain, where learnable tokens are used to condition pre-trained language models without updating their backbone parameters. It was later extended to V-L settings [10–12], where prompts are optimized to better align visual and textual representations for downstream tasks, and further adapted to pure visual models [33, 34] by injecting learnable tokens into the visual encoder to guide feature extraction. Compared with traditional adaptation strategies, prompt learning provides a parameter-efficient alternative by restricting optimization to a small number of additional parameters, thereby preserving the generalization ability of pre-trained models. Existing adaptation approaches typically rely on either full fine-tuning or linear probing [17]. While full fine-tuning offers greater flexibility, it may disrupt the well-established joint visual-linguistic representations learned during pre-training. In contrast, linear probing keeps the backbone frozen but often lacks sufficient capacity to adapt to complex downstream tasks. Moreover, these methods tend to adopt unimodal strategies, optimizing visual and textual components independently, which neglects the inherent multimodal nature of V-L models and limits cross-modal interaction. To address these limitations, we seek to fully exploit the multimodal potential of CLIP, thereby improving the consistency and alignment between visual and linguistic representations for more effective intention understanding in downstream applications.

3.2 Multi-label Intention Understanding

In recent years, with the continuous advancement of deep learning technologies [3–5, 35–41] and their widespread applications in real-world scenarios, research on understanding human intentions has garnered increasing attention. The focus has gradually shifted from early approaches that inferred intentions through observable cues such as facial expressions, body movements, and scene context [42, 43], to more recent efforts that attempt to model the underlying motivations and latent semantics behind human actions [44–46]. This transition reflects a movement from shallow perception to deeper semantic reasoning, where intention understanding is no longer limited to visible patterns but involves inferring implicit and abstract relationships between entities and contexts.

Currently, intention understanding research has been applied across various domains, such as transportation systems [47, 48], where models aim to predict human behaviors for safer interactions, social media platforms [49, 50], where user intentions are analyzed for content recommendation and moderation, and digital advertising [51], where intention cues are leveraged to enhance persuasive strategies. Among pioneering studies in this field, Jia *et al.* [52] constructed the Intentonomy dataset, which explicitly organizes intention categories in a structured manner and provides a foundation for studying the relationships between objects, scene context, and inferred human intent. Building upon such datasets, Shi *et al.* [41, 53, 54] introduced methods that incorporate semantic guidance or label-aware representations to refine intention modeling, enabling the model to better capture high-level semantic dependencies. To further address challenges such as substantial intra-class diversity and ambiguity

between intention categories, Wang *et al.* [55] proposed a prototype-driven framework, where class prototypes are learned to represent central semantic patterns and improve discrimination across categories. Nevertheless, aligning visual representations with textual intention annotations continues to pose a substantial challenge, as the semantic gap between low-level visual features and high-level intention concepts remains difficult to bridge. To mitigate this issue, Yang *et al.* [7] introduced IntCLIP, a vision-language framework that leverages cross-modal alignment to connect visual features with textual intention descriptions.

Despite significant advances in this domain, existing methods still struggle with insufficient alignment between visual and textual modalities during training, particularly when dealing with abstract and ambiguous intention semantics. In addition, reliance on fixed textual categories limits the model's ability to adapt to flexible and evolving intention scenarios, thus constraining its generalization capacity. To address these challenges, we propose a symmetric vision-language framework incorporating dynamic label mechanisms, aiming to establish stronger cross-modal coupling and enhance the adaptability and effectiveness of intention understanding.

3.3 Prompt Learning

Prompt-based adaptation has emerged as an effective paradigm for leveraging the knowledge of pre-trained language models (PLMs) by reformulating downstream tasks into prompt-driven inference. Instead of directly fine-tuning model parameters, this approach conditions the model through carefully designed or learnable prompts, allowing it to reuse pre-trained knowledge in a more controlled manner. This strategy is particularly advantageous in data-scarce scenarios, where it enables models to achieve strong generalization with minimal supervision while avoiding overfitting. With the rapid development of large-scale V-L models, recent efforts have extended prompt learning from pure textual settings to multimodal contexts, aiming to better coordinate visual and linguistic representations within a shared embedding space. Early attempts such as CoOp [11] introduce learnable context tokens to replace manually designed prompts, effectively bridging the gap between pre-trained representations and downstream tasks. Subsequent work further enhances adaptability by incorporating instance-specific visual cues or injecting trainable tokens into the visual encoder, enabling more fine-grained control over feature extraction. More recent approaches [56] explore hierarchical or layer-wise prompt designs, where prompts are inserted at multiple depths of the network to facilitate progressive cross-modal interaction and richer semantic alignment.

Nevertheless, most existing methods primarily focus on refining prompt structures associated with input representations while paying limited attention to how class labels are encoded and utilized within the model. This limitation becomes more pronounced in tasks such as affective or intent recognition, where label semantics are inherently abstract, ambiguous, and often weakly grounded in visual appearance. In such scenarios, the mismatch between fixed textual labels and the learned representation space may hinder effective semantic alignment. In the NLP domain, it has been shown that the choice and formulation of label representations can significantly influence model performance [57], motivating the development of techniques such as automatic

verbalizer construction and continuous label embedding. These methods aim to transform discrete label words into more expressive representations that better capture task-specific semantics. Drawing inspiration from these findings, we move beyond conventional discrete label formulations and instead introduce trainable label embeddings within the multimodal prompt framework. By allowing label representations to be jointly optimized with prompt tokens, this design establishes a more flexible and adaptive mapping between downstream semantic categories and the embedding space of pre-trained V-L models, thereby improving the model's ability to capture subtle semantic distinctions and enhancing robustness in complex recognition scenarios.

4 Proposed Method

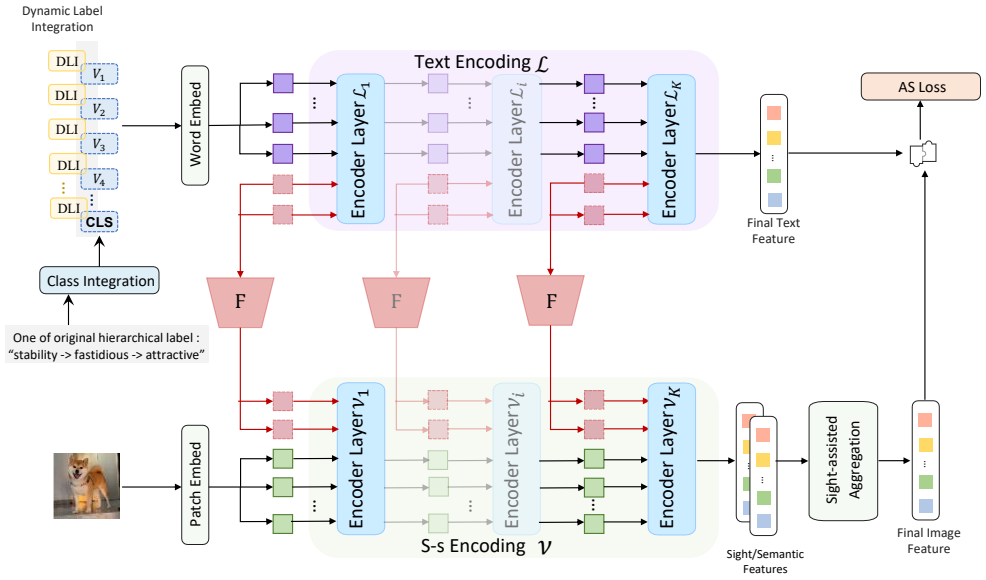


Fig. 5 Overview of our propsed Symmetric Vision-languege Prompting CLIP (SymCLIP). The static class tokens originally defined in the dataset are replaced with trainable embeddings to enhance the generalization capability of the CLIP model in complex semantic spaces. SymCLIP introduces a vision-language joint modeling function $\mathcal{F}$, which enables the generation of vision prompts under the guidance of language prompts, thereby establishing a tighter coupling between the image and text modalities. Finally, an Asymmetric Loss (AS Loss) is employed to optimize the alignment between image and text features, promoting consistent representation of the two modalities within the embedding space.

4.1 Problem Definition

Given an image, the objective of multi-label intention classification is to identify a set of intention-related semantic categories that are present in the visual content. Let

$\mathcal{L} = \{l_1, l_2, \ldots, l_K\}$ denote the complete label space comprising K predefined intention categories. For each training instance x, its ground-truth label vector is represented as $\mathbf{y} = [y_1, y_2, \ldots, y_K]$, where $y_k \in \{1, -1\}$ indicates whether the k-th intention category l_k is relevant (1) or irrelevant (-1) to the image. The goal is to learn a predictive function that identifies the relevant categories for a given image, ensuring that the predicted set of positive labels accurately reflects the underlying intentions conveyed in the visual content.

4.2 Revisiting CLIP

CLIP is a vision-language pre-trained model (VL-PTM) proposed by Radford et al. It consists of an image encoder φ and a text encoder ζ, which are responsible for extracting features from images and texts, respectively, and mapping them to a shared feature space $\mathbb{R}^d$. Based on large-scale contrastive pre-training, CLIP demonstrates superior cross-modal alignment capabilities.

During fine-tuning for intention recognition, we encode each intention category into a semantic vector using a task-specific text prompt (illustrated in Fig. 5), which facilitates explicit correspondence modeling between textual intent representations and visual features. Multi-label classification is then achieved by computing the cosine similarity between the image features and the semantic representations of all categories. However, unlike traditional methods that construct text prompts using fixed templates (e.g., "a photo depicting an {intention class}"), we introduce a Dynamic Label Integration mechanism, which uses a parameterized, learnable vector to replace manually defined sentence structures, allowing the model to automatically adapt to the task semantics and optimize category representations. Additionally, we introduce a Symmetric Vision-Language Prompting mechanism: in the SymCLIP model, learnable tokens are introduced in the first L layers of both the visual and text encoders, establishing a tighter connection between image features and category label prompts.

4.3 Bridging Visual Perception and Semantic Cognition

Modern datasets often adopt hierarchical label structures to represent the intricate semantic relationships among labels. In highly subjective tasks such as MIU, these hierarchies not only enrich label expressiveness but also enhance the model's ability to capture subtle semantic cues. For example, Shi et al.[54] introduced the Hierarchical Label Embedding and Grouping (HLEG) method, which adopts a hierarchical Transformer design to capture dependencies across multiple label levels. However, despite its structural advantages, directly applying hierarchical Transformers to such label systems may lead to redundancy due to overlapping semantics across label levels, resulting in repeated information and increased parameterization, ultimately compromising model efficiency and generalization.

To address this, we adopt a two-stage label integration strategy that leverages large language models and incorporates it into the CLIP prompt learning framework to

improve multimodal alignment. The process is formulated using two prompt templates:

$$\begin{aligned} \text{Prompt}^{+} &= [T_1^{+}, T_2^{+}, \ldots, T_{N^{+}}^{+}, \text{CLS}], \\ \text{Prompt}^{-} &= [T_1^{-}, T_2^{-}, \ldots, T_{N^{-}}^{-}, \text{CLS}], \end{aligned} \tag{3}$$

where $T_j^{\pm}$ are trainable prompt vectors and CLS denotes the class vector produced from hierarchical label integration.

Building upon this, we propose Dynamic Label Integration (DLI) to further align class semantics in downstream tasks with the pretrained model. DLI performs end-to-end optimization of label embeddings and can be seamlessly integrated into intention understanding tasks. The modified prompt template for the i-th class is:

$$h_i = [T_1][T_2] \ldots [T_n][\langle C_i \rangle], \tag{4}$$

where $\langle C_i \rangle$ denotes a trainable class token. The classification probability for class i given an image $\mathcal{I}$ is defined as:

$$P(y = i \mid \mathcal{I}) = \frac{\exp(\cos(g(\mathcal{I}), \phi(h_i))/\gamma)}{\sum_{r=1}^{K} \exp(\cos(g(\mathcal{I}), \phi(h_r))/\gamma)}. \tag{5}$$

To bridge the semantic disparity between visual features and label semantics, our method restricts optimization to the set of class tokens $\langle C_i \rangle_{i=1}^{K}$ during training on each downstream task.

4.4 Symmetric Vision-languege Prompting

In the task of intent classification, previous studies have predominantly relied on unimodal strategies. Shi et al.[53, 54, 58] leveraged semantic information to guide intention modeling, while Wang et al.[55] proposed a prototype-based framework to mitigate intra-class diversity and ambiguity, which often lack the flexibility to dynamically adapt to both linguistic and visual representation spaces. To address this limitation, we explore the potential of multimodal prompt learning and propose a novel method termed Symmetric Vision-Language Prompting (SVLP). This technique emphasizes the structural symmetry and completeness of prompts. Specifically, we introduce learnable tokens in the first L layers of both the vision and language encoders, to progressively model intermediate representations. This hierarchical prompting strategy enables more effective semantic alignment between image and text modalities.

Language Prompting: To enhance the semantic alignment for intent classification, we introduce m trainable context tokens $\{Q_i\}_{i=1}^{m}$ in the textual encoder of CLIP. The initial input embedding sequence is restructured as:

$$[Q_1, Q_2, \ldots, Q_m, T_0], \quad \text{where } T_0 = [t_1, t_2, \ldots, t_L]. \tag{6}$$

Here, T_0 represents the fixed embedding of the intent description tokens. As the input passes through each transformer layer $\mathcal{E}_l$ ($l = 1, 2, \ldots, H$), we append layer-wise

prompts up to a specific depth R, defined as:

$$T_l = \mathcal{E}_l\left([Q_{l-1}, T_{l-1}]\right), \quad l = 1, 2, \ldots, R. \tag{7}$$

Beyond the R-th layer, the remaining transformer layers continue to refine the semantic representations using previously learned prompts:

$$[Q_r, T_r] = \mathcal{E}_r\left([Q_{r-1}, T_{r-1}]\right), \quad r = R+1, \ldots, H. \tag{8}$$

The final intent representation is obtained by projecting the last token embedding after the H-th layer:

$$z^{\text{text}} = \text{Proj}(t_H^L), \tag{9}$$

where t_H^L denotes the embedding of the L-th token at the final layer, and $\text{Proj}(\cdot)$ is the text projection head.

Vision Prompting: Similar to the language prompting strategy, we introduce n learnable context tokens $\{V_i\}_{i=1}^n$ into the vision encoder of CLIP. These tokens are progressively injected as hierarchical prompts across the transformer layers, up to the R-th layer:

$$I_l = \mathcal{V}_l\left([V_{l-1}, E_{l-1}, P_{l-1}]\right), \quad l = 1, 2, \ldots, R, \tag{10}$$

$$[V_r, E_r, P_r] = \mathcal{V}_r\left([V_{r-1}, E_{r-1}, P_{r-1}]\right), \quad r = R+1, \ldots, H, \tag{11}$$

$$z^{\text{image}} = \text{Proj}(p_H). \tag{12}$$

To better exploit the rich pretraining knowledge of CLIP and align it with downstream tasks, we further enhance the CLIP architecture by incorporating a dual-image encoder framework. This design significantly improves the model's ability to capture both explicit visual details and subtle semantic nuances, leading to more comprehensive visual-semantic alignment.

Modality-aware Prompt Fusion: We argue that relying solely on unimodal prompt tuning is insufficient to capture the intricate interactions between visual and linguistic modalities. To achieve comprehensive contextual optimization, it is crucial to adopt a multi-modal co-adaptation strategy that simultaneously aligns both the visual and textual branches of CLIP.

However, existing approaches, such as IntCLIP proposed by Yang et al.[7], typically focus on enabling each modality to learn its task-relevant context independently. Since the visual and language branches operate in isolation during training, such methods often fail to exploit potential synergy between modalities, which is particularly detrimental in tasks like intent recognition where fine-grained alignment is essential.

To address this limitation, we propose a modality-aware prompt fusion mechanism that facilitates joint modeling between language and vision cues. Specifically, language prompt tokens are injected into the first R transformer layers of the text encoder, providing hierarchical semantic guidance. In contrast, visual prompt tokens are no longer independently optimized. Instead, they are dynamically generated from the language prompts via a cross-modal coupling function $\mathcal{F}(\cdot)$, such that:

$$P_h = \mathcal{F}_h(T_h). \tag{13}$$

Algorithm 1 Hierarchical prompting with cross modal coupling

Require: image I, classes K, depth R, maps $F_0, \ldots, F_{R-1}$
Ensure: logits $O \in \mathbb{R}^{B\times 2\times K}$

init context a^+, a^-
init compounds p_ℓ^+, p_ℓ^-, $\ell = 1, \ldots, R-1$
$v_0^+ \leftarrow F_0(a^+)$, $v_0^- \leftarrow F_0(a^-)$
for $\ell = 1$ **to** $R-1$ **do**
    $v_\ell^+ \leftarrow F_\ell(p_\ell^+)$
    $v_\ell^- \leftarrow F_\ell(p_\ell^-)$
end for
for $c = 1$ **to** K **do**
    $s_c^+ \leftarrow [\,\mathrm{sos} \,\|\, a^+ \,\|\, t_c\,]$
    $s_c^- \leftarrow [\,\mathrm{sos} \,\|\, a^- \,\|\, t_c\,]$
    $u_c^+ \leftarrow E_t(s_c^+, \{p_\ell^+\})$
    $u_c^- \leftarrow E_t(s_c^-, \{p_\ell^-\})$
end for
for $b = 1$ **to** B **do**
    $z_b^+ \leftarrow E_v(I_b, a^+, v_0^+, \{v_\ell^+\})$
    $z_b^- \leftarrow E_v(I_b, a^-, v_0^-, \{v_\ell^-\})$
    normalize $z_b^\pm$, $u_c^\pm$
    $\tau \leftarrow \exp(\lambda)$
    for $c = 1$ **to** K **do**
        $O_{b,1,c} \leftarrow \tau\langle z_b^-, u_c^-\rangle$
        $O_{b,2,c} \leftarrow \tau\langle z_b^+, u_c^+\rangle$
    end for
end for

This formulation establishes a dependency of the vision branch on language semantics, enabling the visual pathway to respond adaptively to textual guidance.

$$I_l = \mathcal{V}_l\left([V_{l-1}, E_{l-1}, \mathcal{F}_{l-1}(T_{l-1})]\right), \quad l = 1, 2, \ldots, R, \tag{14}$$

$$[V_r, E_r, P_r] = \mathcal{V}_r\left([V_{r-1}, E_{r-1}, P_{r-1}]\right), \quad r = R+1, \ldots, H, \tag{15}$$

$$z^{\text{image}} = \mathrm{Proj}(p_H). \tag{16}$$

The coupling function $\mathcal{F}_h$ is implemented as a learnable linear transformation that maps linguistic prompt representations into the visual prompt space, serving as a semantic bridge between the two modalities and promoting gradient flow across branches during optimization. The detailed procedure is presented in Algorithm 1.

5 Experiment

5.1 Benchmark setting

Datasets. To investigate the effectiveness of SymCLIP for the multi-intent understanding task (MIU), we conduct experiments on the Intentonomy dataset, which

includes a total of 14,455 images divided into 12,740 for training, 498 for validation, and 1,217 for testing. These images, sourced from Unsplash, are annotated across 28 intent-related categories. SymCLIP establishes explicit links between image prompts and class labels during training, enabling flexible adaptation to diverse intent understanding scenarios.

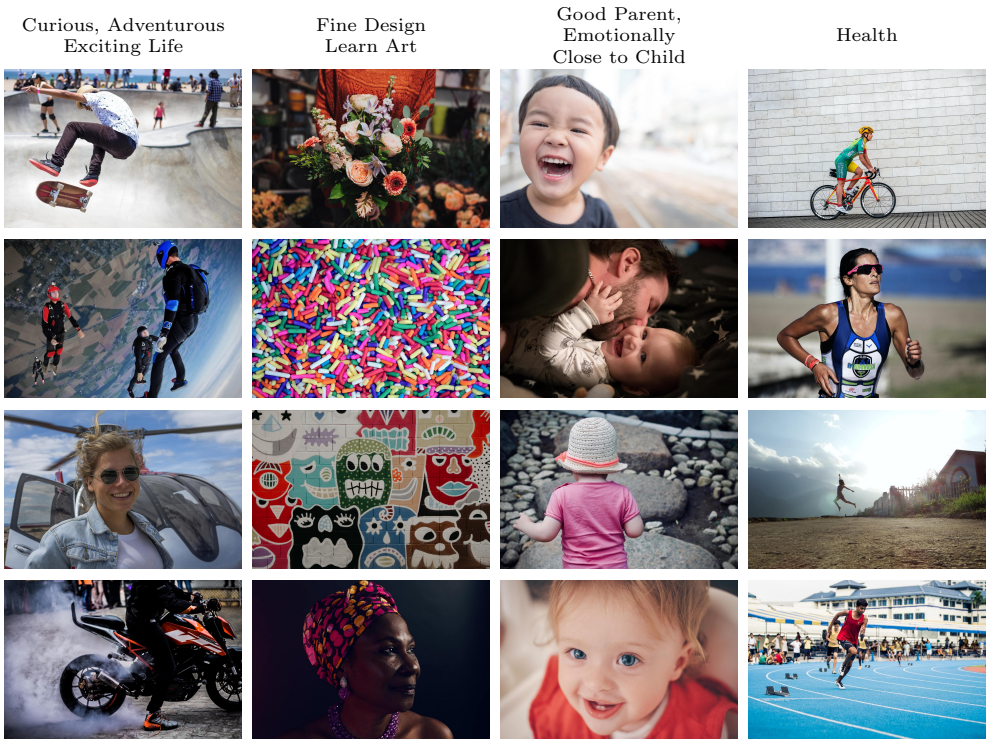


Fig. 6 Image samples of Intentonomy dataset

Performance Assessment Criteria. For MIU evaluation, we employ a composite set of F1-based metrics, including Macro-F1, Micro-F1, and Samples-F1. These metrics provide complementary perspectives for assessing multi-label classification performance.Macro-F1 treats each class equally, regardless of its frequency, offering a balanced view across categories. It is computed as the average of class-wise F1 scores:

$$\text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K}\frac{2 \cdot P_k \cdot R_k}{P_k + R_k}, \tag{17}$$

where K denotes the total number of classes, and P_k and R_k represent the precision and recall of the k-th class, respectively.

Micro-F1 aggregates contributions from all classes to compute a global performance score, which is particularly suitable for imbalanced datasets. It is defined as:

$$\text{Micro-F1} = \frac{2 \cdot \sum_{k=1}^{K} TP_k}{2 \cdot \sum_{k=1}^{K} TP_k + \sum_{k=1}^{K} FP_k + \sum_{k=1}^{K} FN_k}, \tag{18}$$

where TP_k, FP_k, and FN_k denote the true positives, false positives, and false negatives for class k, respectively.Samples-F1 evaluates the model at the instance level, reflecting its behavior on individual examples. It is computed by averaging the F1 score over all samples:

$$\text{Samples-F1} = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \cdot |Y_i \cap \hat{Y}_i|}{|Y_i| + |\hat{Y}_i|}, \tag{19}$$

where N is the number of samples, Y_i denotes the ground-truth label set, and $\hat{Y}_i$ represents the predicted label set for the i-th instance.

The average of these three metrics provides a holistic assessment of model capability from class-level, global-level, and instance-level perspectives.In addition, we adopt mean Average Precision (mAP) as a supplementary metric to evaluate the ranking quality of predicted labels. It is defined as:

$$\text{mAP} = \frac{1}{K} \sum_{k=1}^{K} AP_k, \tag{20}$$

where AP_k denotes the average precision of the k-th class.

Implementation Details. To ensure consistency with previous vision-language models, all experiments are conducted on an NVIDIA RTX 3090 GPU, and we follow the settings in[56], both visual and textual encoders are initialized from the CLIP ViT-B/16 model, where the textual embedding dimension is set to $d_l = 512$, the visual token dimension is $d_v = 768$, and the multimodal fused prompt dimension is $d_{vl} = 512$, with input images standardized at a resolution of 224×224 pixels. The Transformer-based text encoder and the visual encoder's sight branch remain fixed throughout the training phase to preserve pre-trained representational priors. Inspired by [11], we associate each semantic label with two context embeddings, each comprising 16 learnable tokens. Furthermore, we explicitly define the prompt depth as $J = 9$, and set both the linguistic and visual prompt lengths to 2, allowing for more structured prompt tuning. Optimization is carried out using SGD, initialized at a learning rate of 0.002, following a cosine annealing decay strategy. The model is trained for 50 epochs, employing batch sizes of 32 for intent recognition tasks. Following prior work [2], we adopt asymmetric loss with parameters $\gamma^+ = 1$, $\gamma^- = 2$, and margin $c = 0.05$ to better accommodate the class imbalance inherent in multi-label learning settings.

Baseline Setup. To comprehensively evaluate the adaptability and performance of the proposed SymCLIP model in the MIU task, we establish a benchmark suite that spans three representative methodological categories.

For the MIU task, we compare against several specialized models that were originally designed for this setting, including CPAD [58], the baseline method in

Table 1 We conduct a comprehensive performance comparison of various state-of-the-art multi-label classification and intention understanding methods on the Intentonomy dataset, evaluated on both the testing set (2K images) and the validation set (2K images). Among the evaluated approaches, our proposed **SymCLIP** model, built upon the Vision Transformer backbone (ViT-B/16), consistently achieves superior results across all key metrics.

Methods	Intentonomy Testing(2K)				Intentonomy Validation(2K)			
	MacroF1	MicroF1	SamplesF1	mAP	MacroF1	MicroF1	SamplesF1	mAP
Multi-label classification methods								
MultiGuard (Jia et al., 2022) [60]	26.87	38.37	40.06	36.08	27.74	41.03	41.58	37.55
STT (Chen et al., 2022) [61]	27.33	42.21	42.62	35.93	28.78	41.03	41.75	32.12
Query2Label (Liu et al., 2021) [59]	32.12	44.64	45.15	34.33	35.35	46.34	47.40	38.86
DualcoOp (Sun et al., 2022) [2]	31.06	44.42	40.40	39.33	32.67	46.29	42.97	40.93
Intention Understanding methods								
Intentonomy (Jia et al., 2021) [52]	23.98	31.28	31.39	-	25.07	32.94	33.61	-
CPAD (Ye et al., 2023) [58]	27.39	40.98	41.12	-	27.37	41.77	42.68	-
HLEG (Shi et al., 2023) [54]	32.77	44.69	45.54	35.93	35.35	46.34	47.40	36.86
PIP-Net (Wang et al., 2023) [55]	32.57	45.94	47.05	-	30.71	44.84	45.48	-
LabCR (Shi et al., 2024) [53]	34.63	48.51	48.05	37.13	37.10	49.04	49.71	38.86
IntCLIP (R101) (Yang et al., 2024) [7]	38.40	50.54	49.31	42.66	40.09	50.68	48.99	42.84
IntCLIP (ViT-B/32) (Yang et al., 2024) [7]	40.15	53.54	51.01	45.32	41.55	52.61	49.75	44.95
Ours	**45.06**	**55.38**	**54.70**	**49.26**	**43.67**	**53.36**	**51.54**	**51.55**

Intentonomy [52], PIP-Net [55], and HLEG [54]. These methods are re-implemented and evaluated under the same dataset and evaluation protocol as originally proposed in [54, 55] .

Considering the strong relevance between MIU and multi-label classification, we additionally incorporate a set of competitive multi-label learning baselines: Query2Label [59], MultiGuard [60], DualCoOp [2], and SST [61]. Notably, DualCoOp leverages large-scale pre-trained vision-language models with a dual-context prompting strategy, which significantly improves its generalization capacity in multi-label scenarios.

Finally, to enable cross-task validation, we evaluate our model against several prominent methods in the domain of image emotion recognition, including MSRCA [62], ECWA [63], CGLF-Net [64], and WSEIP [65]. Unlike these methods, SymCLIP employs a unified prompt-based vision-language framework, enabling consistent representation learning across both intent and emotion understanding tasks.

5.2 Multi-Label Intention Understanding

To evaluate the performance of our proposed SymCLIP model in intention recognition, we conducted a comprehensive comparative study on the Intentonomy dataset, which consists of a testing set and a validation set. A diverse set of representative and state-of-the-art methods were selected for comparison, and the results are summarized in Table 1. For clarity, these baselines are divided into two categories: general-purpose multi-label classification methods and approaches specifically designed for multi-label intention understanding.

From the results, it can be observed that conventional multi-label classification methods, such as MultiGuard[60], STT[61], and Query2Label[59], exhibit relatively limited performance on this task. This is mainly because they lack explicit modeling of high-level semantic intentions and are insufficient in capturing the complex relationships between visual content and abstract intent labels. In contrast, intention-oriented methods, including HLEG[54], PIP-Net[55], LabCR[53], and IntCLIP, achieve

Table 2 A comparative analysis of different learnable parameter sizes reveals that the model achieves optimal performance when the parameter size is set to 512.

Learnable params	Intentonomy Testing (2K)				Intentonomy Validation (2K)			
	MacroF1	MicroF1	SamplesF1	mAP	MacroF1	MicroF1	SamplesF1	mAP
64	41.85	51.83	49.06	45.47	41.85	52.34	50.45	47.96
128	41.57	52.78	51.05	47.76	41.61	53.32	51.89	48.79
256	42.18	52.98	51.44	46.84	**44.33**	**54.60**	**53.32**	50.46
512	**45.06**	**55.38**	**54.70**	**49.26**	43.67	53.36	51.54	**51.55**
1024	42.13	53.29	52.06	46.77	42.53	54.63	54.04	49.24

notable improvements by incorporating semantic structures, prototype learning, or vision–language alignment mechanisms.

Despite these advancements, our proposed SymCLIP consistently outperforms all competing methods across all evaluation metrics on both the testing and validation sets. Specifically, on the Intentonomy testing set, SymCLIP achieves a Micro-F1 score of 55.38%, surpassing the strongest baseline, IntCLIP[7] (ViT-B/32), by a margin of 1.84%. More importantly, SymCLIP also demonstrates clear advantages in Macro-F1 (45.06%) and Samples-F1 (54.70%), indicating improved performance not only on dominant classes but also on long-tail and instance-level predictions. In terms of mAP, our method reaches 49.26%, which further confirms its superior ranking capability in multi-label prediction scenarios.

Consistent improvements are also observed on the validation set, where SymCLIP achieves 53.36% in Micro-F1 and 51.55% in mAP, consistently outperforming all baseline methods. The stable improvements across both data splits demonstrate the strong generalization ability of our framework.

Performance gains can be attributed to the complementary design of the two key components in SymCLIP. On the one hand, SVLP introduces hierarchical prompt tokens into both visual and textual encoders, enabling progressive cross-modal alignment and allowing visual representations to be explicitly guided by semantic cues. On the other hand, DLI replaces fixed label words with trainable embeddings, which adaptively refine class representations and reduce semantic ambiguity among intent categories. Joint optimization of feature-level alignment and label-level representation effectively bridges the gap between visual perception and semantic understanding, leading to more accurate and robust intention recognition.

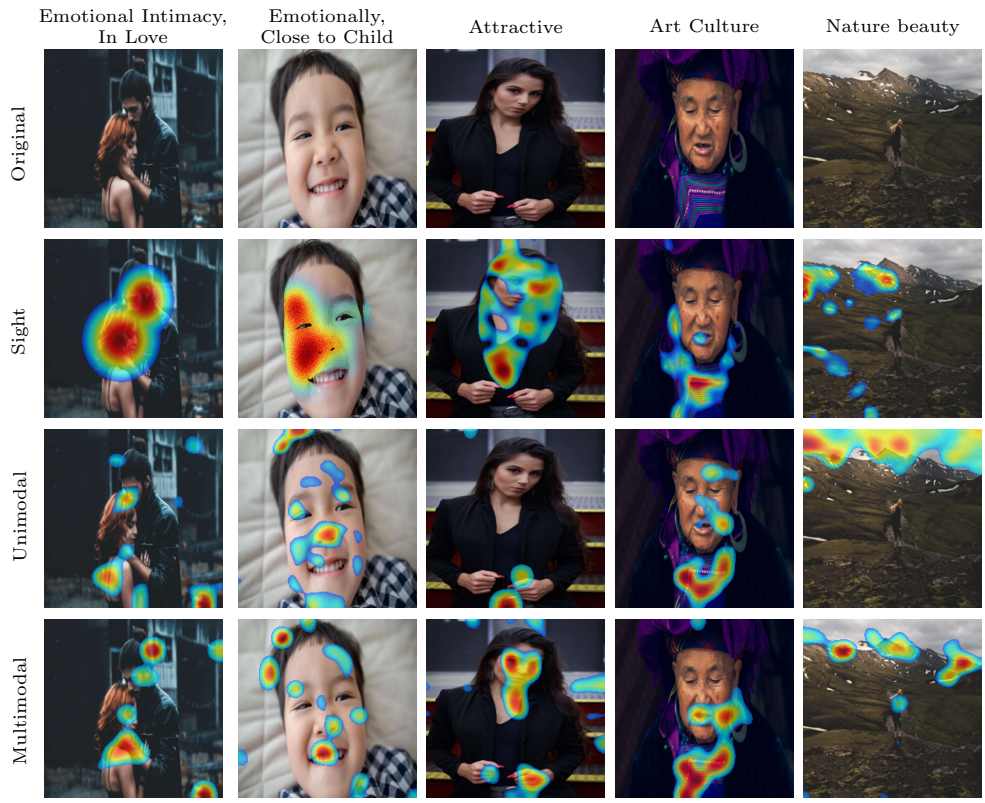


Fig. 7 SVLP attention visualization. The first row shows original images with intent labels, followed by attention maps from the sight encoder, a unimodal optimization baseline, and our SVLP model. Our method exhibits more accurate and concentrated responses on intent-relevant regions, demonstrating the effectiveness of cross-modal coordination over unimodal optimization.

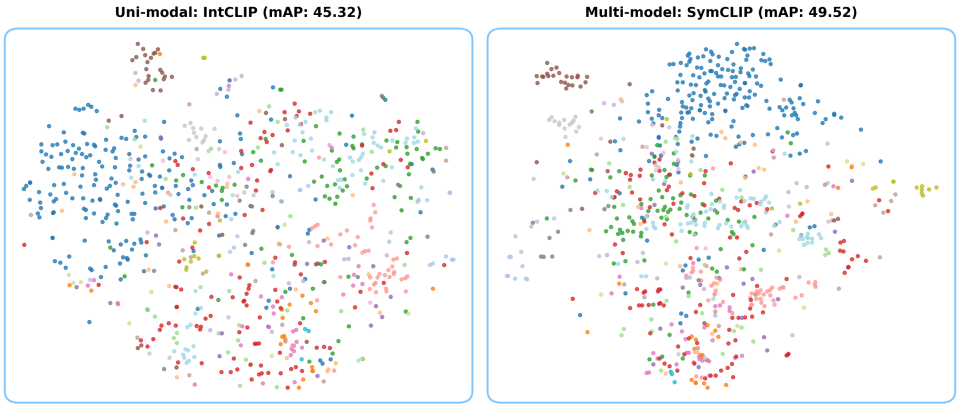


Fig. 8 t-SNE visualization of image embeddings on the Intentonomy dataset for IntCLIP (unimodal prompting) and the proposed SymCLIP. The results show that SymCLIP achieves improved separability across intent categories.

Table 3 A comparative analysis demonstrates that enabling both SVLP and DLI yields the best performance.

SVLP	DLI	Intentonomy Testing (2K)				Intentonomy Validation (2K)			
		SamplesF1	MacroF1	MicroF1	mAP	SamplesF1	MacroF1	MicroF1	mAP
✓		52.45	43.28	52.08	46.16	50.39	41.09	51.02	47.69
	✓	50.98	41.59	51.78	47.34	49.08	40.48	50.89	48.57
✓	✓	**54.70**	**45.06**	**55.38**	**49.26**	**51.54**	**43.67**	**53.36**	**51.55**

5.3 Ablation Experiments

We conducted a comprehensive set of ablation studies and visualization analyzes on the Intentonomy dataset to validate the effectiveness of the proposed SymCLIP framework, with particular emphasis on the contributions of the SVLP and DLI modules. In addition, we investigate the impact of key design choices, including prompt structure and learnable parameter settings, to provide a deeper understanding of the model behavior. In addition, qualitative visualization results are presented to illustrate the discriminative capacity of learned representations. Finally, we analyze the limitations of the proposed method to offer insights for future improvements.

5.3.1 Hyperparameter Sensitivity Analysis

We investigate the effect of varying the number of learnable parameters on model performance using the Intentonomy dataset. As shown in Table 2, increasing the number of learnable parameters generally leads to improved performance on the Intentonomy Testing (2K) set, indicating that a larger parameter space enables the model to learn more expressive and discriminative representations.

Specifically, when the number of learnable parameters increases from 64 to 512, all evaluation metrics exhibit a consistent upward trend. The best performsance is achieved at 512 parameters, where the model obtains a Micro-F1 score of 55.38%, a Samples-F1 score of 54.70%, and the highest mAP of 49.26%. This suggests that an appropriate number of learnable parameters allows the model to effectively capture complex intent semantics while maintaining stable optimization. In particular, the improvement in Macro-F1 further indicates enhanced capability in handling class imbalance and learning more balanced representations across categories. However, when the parameter size is further increased to 1024, the performance on several metrics declines, with Micro-F1 dropping to 53.29% and mAP decreasing to 46.77%. This degradation implies that excessive parameterization may introduce redundancy and increase the risk of overfitting, thereby weakening the model's generalization ability.

A similar trend is observed on the Intentonomy Validation (2K) set. Configurations with 256 and 512 parameters achieve relatively strong and stable performance. Notably, the 256-parameter setting attains the highest Macro-F1 (44.33%) and Micro-F1 (54.60%), indicating its effectiveness in generalization. Meanwhile, the 512-parameter configuration achieves the best mAP of 51.55%, reflecting superior ranking capability across multiple labels. Although the 1024-parameter setting achieves a slightly higher Micro-F1 of 54.63%, its Macro-F1 and mAP are inferior to those of the 512-parameter configuration, suggesting that the additional parameters do not translate into consistent performance gains.

5.3.2 Effectiveness Analysis of Key Components

Our framework integrates two key components, namely SVLP and DLI. To systematically evaluate their individual contributions and joint effectiveness, we conducted ablation studies as summarized in Table 3. As observed, each module independently brings consistent performance gains across all evaluation metrics, validating their effectiveness in improving multimodal representation learning for intention understanding.

Specifically, when only SVLP is enabled, the model achieves a Micro-F1 score of 52.08% on the Intentonomy Testing (2K) set. This improvement can be attributed to the symmetric prompt design, where hierarchical prompt tokens are introduced into both the visual and textual encoders. By establishing a structured interaction between the two modalities, SVLP enables progressive cross-modal alignment and allows visual representations to be explicitly guided by semantic cues from the language branch. This mechanism alleviates the limitation of conventional unimodal prompt tuning and enhances the model's ability to capture implicit and high-level intentions.

In contrast, employing only DLI yields a Micro-F1 score of 51.78% and a relatively higher mAP of 47.34%. The performance gain mainly stems from the adaptive optimization of class representations. Instead of relying on fixed textual templates, DLI introduces trainable class tokens that are optimized in an end-to-end manner. This design allows the model to dynamically adjust label embeddings according to the data distribution, thereby reducing semantic ambiguity among intention categories and improving fine-grained discrimination capability.

When both SVLP and DLI are jointly applied, the model achieves the best performance across all metrics, reaching 55.38% in Micro-F1, 54.70% in Samples-F1, and the highest mAP of 49.26% on the testing set. This significant improvement demonstrates a strong synergistic effect between the two components. From a mechanistic perspective, SVLP enhances feature-level alignment by coupling visual and textual prompts through hierarchical and cross-modal interactions, while DLI refines label-level semantics by enabling flexible and learnable category representations. The joint optimization of feature space and label space effectively bridges the gap between visual perception and semantic cognition, leading to more accurate and robust intention recognition.

A similar trend is observed on the Intentonomy Validation (2K) set, where the combined configuration achieves a Micro-F1 score of 53.36%, outperforming the configurations with only SVLP (51.02%) or only DLI (50.89%). Consistent improvements across Macro-F1, Samples-F1, and mAP further confirm the stability and generalization ability of the proposed framework under different data splits.

To further assess the effectiveness of the proposed multimodal learning mechanism, we present Grad-CAM(Fig. 7) visualizations that illustrate how different strategies attend to semantically meaningful regions. The proposed method consistently produces more concentrated and informative attention. In scenes involving interpersonal relationships, it accurately focuses on critical cues such as faces and hand interactions, which are essential for conveying emotional intimacy. Similarly, in images of children and adult portraits, the model demonstrates refined sensitivity to facial structures, capturing key features more precisely and enabling a better interpretation of

subtle emotional expressions. In culturally rich scenarios, it further highlights distinctive clothing patterns and visual details, facilitating a more reliable understanding of cultural and artistic intent.

In contrast, unimodal approaches tend to generate scattered or less relevant attention, especially in complex scenes. For instance, attention may drift toward irrelevant regions while overlooking semantically important elements. This limitation is particularly evident in natural landscapes, where unimodal methods fail to adequately capture key structures such as mountain ranges. The proposed cross-modal framework, guided by language semantics, instead attends to more meaningful and comprehensive regions, leading to a more holistic and accurate understanding of the scene. These results demonstrate the superiority of cross-modal coordination in enhancing both localization precision and high-level semantic reasoning.

Moreover, the embedding visualization(Fig. 8) demonstrates that our method yields a significantly more structured feature space, characterized by enhanced intra-class compactness and pronounced inter-class separability. In contrast to existing approaches, which exhibit noticeable overlap among semantic categories, the proposed model produces well isolated clusters with clear decision boundaries. This suggests that the learned representations are not only more discriminative but also more semantically aligned.

Overall, the ablation results indicate that SVLP and DLI are highly complementary. SVLP primarily focuses on enhancing cross-modal feature interaction, while DLI targets semantic label adaptability. Their integration enables coordinated optimization in both representation learning and semantic modeling, which is crucial for addressing the inherent ambiguity and diversity in multi-label intention understanding tasks. The consistent superiority of the combined configuration verifies the necessity of incorporating both components within the proposed architecture.

5.3.3 Limitations

Although the proposed SymCLIP framework achieves competitive performance across multiple benchmarks, several limitations remain due to the reliance on pre-trained representations, dataset characteristics, and modeling assumptions. A systematic analysis of these aspects not only provides deeper insight into the behavior of the model but also suggests concrete directions for future improvements. The suboptimal scenarios can be summarized as follows:

Dependence on pre-trained vision–language alignment. SymCLIP builds on the pre-trained CLIP model, and its effectiveness is therefore closely tied to the quality of the initial vision–language alignment. When the downstream task exhibits domain-specific intent patterns that differ significantly from the pre-training distribution, the learned representations may become suboptimal. This issue is particularly evident in fine-grained scenarios where subtle semantic distinctions are required. Although the proposed prompt-based adaptation mitigates this limitation to some extent by refining task-specific representations, it still fundamentally depends on the representational capacity of the pre-trained backbone, which may constrain performance under substantial domain shifts.

Ambiguity in high-level semantic interpretation. Image-based intention and emotion recognition inherently involve subjective and high-level semantic understanding, where different categories may share subtle visual patterns. In cases where visual cues are weak or semantically overlapping, the model may produce confused predictions across related categories. This limitation reflects the challenge of aligning abstract semantic concepts with visual representations, especially when the semantic boundaries between categories are not well defined.

Insufficient representation of rare or long-tail categories. The distribution of intention and emotion categories is often highly imbalanced. Due to limited training samples for rare classes, the model may fail to capture their distinctive characteristics, resulting in unstable predictions. While the Dynamic Label Integration (DLI) module enhances label adaptability through learnable representations, its effectiveness is still constrained when data for certain categories is extremely scarce.

Dependency on structured label semantics. The effectiveness of the DLI module is partially based on the availability of structured or semantically meaningful label relationships, as exemplified by the Intentonomy dataset. In such settings, hierarchical or correlated label structures can be leveraged to improve representation learning. However, when the model is applied to datasets with flat or weakly structured label spaces, where explicit relationships between categories are absent, the benefit of modeling label interactions may be diminished. Consequently, this dependency may restrict the general applicability of the framework to more diverse or unstructured scenarios.

The above limitations inevitably affect the overall accuracy of the prediction. Several potential directions may help alleviate these issues. Expanding the evaluation to more diverse datasets, especially those with complex semantic structures or limited resources, would provide a more comprehensive assessment. Additionally, improving dataset quality and label annotations, as well as incorporating more advanced cross-modal modeling strategies, could further enhance robustness. These directions offer promising avenues for future research in multimodal intention and emotion understanding.

5.4 Cross-Task Generalization Experiments

To evaluate the generalization capability of the proposed method in subjective perception tasks, we extend its application to image-based emotion recognition (IER), a representative domain that inherently involves strong subjectivity and abstract semantic interpretation. Compared with intention understanding, emotion recognition further requires the model to capture subtle affective cues and implicit semantic associations, posing a more challenging test for generalization ability.

For this purpose, we conduct empirical evaluations on two widely recognized emotion recognition datasets, EmotionROI_6 and FI_8. The experimental results, as presented in Table 4, demonstrate that SymCLIP consistently outperforms existing state-of-the-art methods across all metrics. In particular, our method achieves 72.09% on Emo_6 and 80.07% on FI_8, with an average accuracy of 76.08%, surpassing the strong baseline IntCLIP by +1.04%, +1.14%, and +1.09%, respectively.

Table 4 Generalization analysis of IntPrompt on Image Emotion Recognition (IER) tasks.

Methods	Emo_6	FI_8	Avg.
ECWA (Deng et al.) [63]	59.09	70.87	64.98
WSEIP (Zhang et al.) [65]	60.41	75.91	68.16
MSRCA (Zhang et al.) [62]	55.60	72.60	64.10
CGLF-Net (Luo et al.) [64]	65.01	75.61	70.31
IntCLIP (Yang et al.) [7]	71.05	78.93	74.99
Ours	**72.09**	**80.07**	**76.08**
Δ *Improvement vs. IntCLIP*	+1.04	+1.14	+1.09

These improvements, although moderate, are consistent across datasets, highlighting the robustness of the proposed framework.

The observed performance improvements further indicate that the proposed framework maintains strong generalization capability when transferred to subjective perception tasks beyond intention understanding. Despite the inherent differences between intention recognition and emotion recognition, SymCLIP demonstrates stable and consistent gains across both datasets, suggesting that the learned representations are not overly specialized to a single task. Instead, the model is able to capture transferable semantic patterns that generalize well to domains involving abstract and affective reasoning.

6 Conclusion

Multi-label Intent Understanding (MIU) in social media imagery presents significant challenges due to the abstract, implicit, and often multimodal nature of intent, which frequently extends beyond observable visual elements. To tackle these complexities, we introduce SymCLIP, a novel framework designed to address the uncertainty and diversity inherent in intent interpretation. SymCLIP incorporates two core components: a symmetric vision-language prompting mechanism and a dynamic label integration strategy, both of which collaboratively enhance semantic modeling in intent understanding tasks.

In contrast to previous approaches, SymCLIP emphasizes robust cross-modal interaction by establishing mutually reinforcing alignments between visual and textual representations. Empirical evaluations demonstrate that SymCLIP consistently surpasses state-of-the-art methods not only on standard MIU benchmarks but also on more subjective and complex tasks such as image-based emotion recognition. By leveraging large-scale multimodal pretraining, SymCLIP enables fine-grained intent interpretation in user-generated visual content, offering novel insights into computational modeling of semantic understanding in digital communication.

Limitations and Future Work Despite its effectiveness, SymCLIP still has several limitations. The reliance on cross-modal prompting and learnable label representations introduces additional computational overhead, and the model may face

challenges in handling highly ambiguous or overlapping intent categories commonly found in social media content. In addition, its performance depends on the quality of multimodal pretraining, which may limit generalization under domain changes. Addressing these issues remains an important direction for future work.

Multi-task Learning and Multimodal Extension Building upon the current framework, several promising directions can be further explored. First, the integration of multi-task learning offers promising potential. Jointly modeling intention understanding with related tasks, such as emotion recognition, sentiment analysis, or visual reasoning, may further enhance semantic consistency and improve overall performance through shared representations. Second, richer multimodal fusion strategies could be explored by incorporating additional modalities, such as audio, temporal signals, or textual context from social media, to provide more comprehensive cues for intent inference. Advanced cross-modal alignment techniques, including attention-based fusion or generative modeling, may further strengthen the interaction between modalities.

Robustness and Future Data Development In addition, improving robustness under real-world conditions remains an important challenge. Social media imagery often exhibits high variability in terms of scene complexity, noise, and user-specific expression styles. Future work may explore data augmentation strategies, domain generalization techniques, and large-scale pretraining to enhance model adaptability across diverse distributions. Finally, the development of high-quality, fine-grained MIU datasets is crucial. Constructing datasets with richer annotations, diverse intent categories, and varying levels of semantic abstraction would provide a stronger foundation for advancing research in this field.

Declarations

Author contributions S.W. wrote the main manuscript text and H.M. prepared all figures and tables. All authors have read and agreed to the published version of the manuscript.
Funding The authors declare that no funding was received for this work.
Data availability The datasets used in this study are publicly available from their respective official sources.
Competing interests The authors declare no competing interests.
Ethical approval Ethical approval was not required for this study.
Consent for publication All authors approved the final manuscript for publication.
Consent for participation All authors agreed to participate in this study.

References

[1] 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 **11**, 100103 (2021)

[2] Sun, X., Hu, P., Saenko, K.: Dualcoop: Fast adaptation to multi-label recognition with limited annotations. Advances in Neural Information Processing Systems **35**, 30569–30582 (2022)

[3] Chen, S., Ye, M., Du, B.: Rotation invariant transformer for recognizing object in uavs. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 2565–2574 (2022)

[4] Li, C., Chen, S., Ye, M.: Adaptive high-frequency transformer for diverse wildlife re-identification. In: European Conference on Computer Vision, pp. 296–313 (2024). Springer

[5] 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 **46**(2), 712–728 (2023)

[6] 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: International Conference on Machine Learning, pp. 8748–8763 (2021). PmLR

[7] Yang, Q., Ye, M., Tao, D.: Synergy of sight and semantics: visual intention understanding with clip. In: European Conference on Computer Vision, pp. 144–160 (2024). Springer

[8] Wang, Z., Zhang, Z., Lee, C.-Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 139–149 (2022)

[9] Zhang, Y., Zhou, K., Liu, Z.: Neural prompt search. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

[10] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16816–16825 (2022)

[11] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)

[12] Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15659–15669 (2023)

[13] 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: International Conference on Machine Learning,

pp. 4904–4916 (2021). PMLR

[14] Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18123–18133 (2022)

[15] Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)

[16] Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)

[17] 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 **132**(2), 581–595 (2024)

[18] Kim, K., Laskin, M., Mordatch, I., Pathak, D.: How to adapt your large-scale vision-and-language model (2021)

[19] 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 preprint arXiv:2111.03930 (2021)

[20] Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images. In: European Conference on Computer Vision, pp. 701–717 (2022). Springer

[21] Gu, X., Lin, T.-Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921 (2021)

[22] Maaz, M., Rasheed, H., Khan, S., Khan, F.S., Anwer, R.M., Yang, M.-H.: Class-agnostic object detection with multi-modal transformer. In: European Conference on Computer Vision, pp. 512–531 (2022). Springer

[23] Bangalath, H., Maaz, M., Khattak, M.U., Khan, S.H., Shahbaz Khan, F.: Bridging the gap between object and image-level representations for open-vocabulary detection. Advances in Neural Information Processing Systems **35**, 33781–33794 (2022)

[24] Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-vocabulary detr with conditional matching. In: European Conference on Computer Vision, pp. 106–122 (2022). Springer

[25] Zhou, X., Girdhar, R., Joulin, A., Krähenbühl, P., Misra, I.: Detecting twenty-thousand classes using image-level supervision. In: European Conference on Computer Vision, pp. 350–368 (2022). Springer

[26] Ding, J., Xue, N., Xia, G.-S., Dai, D.: Decoupling zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11583–11592 (2022)

[27] Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022)

[28] Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7086–7096 (2022)

[29] Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18082–18091 (2022)

[30] Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691 (2021)

[31] Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190 (2021)

[32] Liu, X., Ji, K., Fu, Y., Tam, W.L., Du, Z., Yang, Z., Tang, J.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602 (2021)

[33] Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.-N.: Visual prompt tuning. In: European Conference on Computer Vision, pp. 709–727 (2022). Springer

[34] Wang, Z., Zhang, Z., Ebrahimi, S., Sun, R., Zhang, H., Lee, C.-Y., Ren, X., Su, G., Perot, V., Dy, J., *et al.*: Dualprompt: Complementary prompting for rehearsal-free continual learning. In: European Conference on Computer Vision, pp. 631–648 (2022). Springer

[35] Huang, W., Ye, M., Shi, Z., Li, H., Du, B.: Rethinking federated learning with domain shift: A prototype view. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16312–16322 (2023). IEEE

[36] Shi, W., Ye, M.: Prototype reminiscence and augmented asymmetric knowledge aggregation for non-exemplar class-incremental learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1772–1781 (2023)

[37] Wu, J., Zheng, R., Jiang, J., Tian, Z., Chen, W., Wang, Z., Yu, F.R., Leung, V.C.: A lightweight small object detection method based on multi-layer coordination federated intelligence for coal mine iovt. IEEE Internet of Things Journal (2024)

[38] 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 **32**, 4543–4554 (2023)

[39] Yang, Q., Ye, M., Du, B.: Emollm: Multimodal emotional understanding meets large language models. arXiv preprint arXiv:2406.16442 (2024)

[40] Ye, M., Chen, S., Li, C., Zheng, W.-S., Crandall, D., Du, B.: Transformer for object re-identification: A survey. International Journal of Computer Vision, 1–31 (2024)

[41] Ye, M., Wu, Z., Chen, C., Du, B.: Channel augmentation for visible-infrared re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(4), 2299–2315 (2023)

[42] Joo, J., Li, W., Steen, F.F., Zhu, S.-C.: Visual persuasion: Inferring communicative intents of images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 216–223 (2014)

[43] Joo, J., Steen, F.F., Zhu, S.-C.: Automated facial trait judgment and election outcome prediction: Social dimensions of face. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3712–3720 (2015)

[44] 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 **35**(12), 2891–2903 (2013)

[45] Vondrick, C., Oktay, D., Pirsiavash, H., Torralba, A.: Predicting motivations of actions by leveraging text. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2997–3005 (2016)

[46] 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. arXiv preprint arXiv:1906.06725 (2019)

[47] 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 **5**(2), 3485–3492 (2020)

[48] Zyner, A., Worrall, S., Ward, J., Nebot, E.: Long short term memory for driver intent prediction. In: 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 1484–1489 (2017). IEEE

[49] Ghaisani, A.P., Handayani, P.W., Munajat, Q.: Users' motivation in sharing information on social media. Procedia Computer Science **124**, 530–535 (2017)

[50] Stavros, C., Meng, M.D., Westberg, K., Farrelly, F.: Understanding fan motivation for interacting on social media. Sport management review **17**(4), 455–469 (2014)

[51] 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 (ANZMAC) (2012). Australian and New Zealand Marketing Academy

[52] Jia, M., Wu, Z., Reiter, A., Cardie, C., Belongie, S., Lim, S.-N.: Intentonomy: a dataset and study towards human intent understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12986–12996 (2021)

[53] 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)

[54] Shi, Q., Ye, M., Zhang, Z., Du, B.: Learnable hierarchical label embedding and grouping for visual intention understanding. IEEE Transactions on Affective Computing **14**(4), 3218–3230 (2023)

[55] Wang, B., Yang, K., Zhao, Y., Long, T., Li, X.: Prototype-based intent perception. IEEE Transactions on Multimedia **25**, 8308–8319 (2023)

[56] Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19113–19122 (2023)

[57] Schick, T., Schütze, H.: Exploiting cloze-questions for few-shot text classification and natural language inference. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 255–269 (2021)

[58] Ye, M., Shi, Q., Su, K., Du, B.: Cross-modality pyramid alignment for visual intention understanding. IEEE Transactions on Image Processing **32**, 2190–2201 (2023)

[59] Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Query2label: A simple transformer way to multi-label classification. arXiv preprint arXiv:2107.10834 (2021)

[60] Jia, J., Qu, W., Gong, N.: Multiguard: Provably robust multi-label classification against adversarial examples. Advances in Neural Information Processing Systems **35**, 10150–10163 (2022)

[61] Chen, T., Pu, T., Wu, H., Xie, Y., Lin, L.: Structured semantic transfer for multi-label recognition with partial labels. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 339–346 (2022)

[62] Zhang, J., Liu, X., Chen, M., Ye, Q., Wang, Z.: Image sentiment classification via multi-level sentiment region correlation analysis. Neurocomputing **469**, 221–233 (2022)

[63] 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, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part I 1, pp. 553–564 (2021). Springer

[64] 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 **26**, 1894–1908 (2023)

[65] Zhang, H., Xu, M.: Weakly supervised emotion intensity prediction for recognition of emotions in images. IEEE Transactions on Multimedia **23**, 2033–2044 (2020)