Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

GMCoT: a graph-augmented multimodal chain-of-thought reasoning framework for multi-label zero-shot learning*#

Xiang WEN†1, Haobo WANG3, Ke CHEN1,2, Tianlei HU1,2, Gang CHEN†‡1,2

[1]State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou 310027, China

[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security, Zhejiang University, Hangzhou 310027, China

[3]School of Software Technology, Zhejiang University, Hangzhou 310027, China

†E-mail: wenxiang@zju.edu.cn; cg@zju.edu.cn

Received June 23, 2025; Revision accepted Dec. 2, 2025; Crosschecked Dec. 11, 2025

Abstract: In recent years, multi-label zero-shot learning (ML-ZSL) has garnered increasing attention because of its wide range of potential applications, such as image annotation, text classification, and bioinformatics. The central challenge in ML-ZSL lies in predicting multiple labels for unseen classes without requiring any labeled training data, which contrasts with conventional supervised learning paradigms. However, existing methods face several significant challenges. These include the substantial semantic gap between different modalities, which impedes effective knowledge transfer, and the intricate and typically complex relationships among multiple labels, making it difficult to model them in a meaningful and accurate manner. To overcome these challenges, we propose a graph-augmented multimodal chain-of-thought (GMCoT) reasoning approach. The proposed method combines the strengths of multimodal large language models with graph-based structures, significantly enhancing the reasoning process involved in multi-label prediction. First, a novel multimodal chain-of-thought reasoning framework is presented which imitates human-like step-by-step reasoning to produce multi-label predictions. Second, a technique is presented for integrating label graphs into the reasoning process. This technique enables the capture of complex semantic relationships among labels, thereby improving the accuracy and consistency of multi-label generation. Comprehensive experiments on benchmark datasets demonstrate that the proposed GMCoT approach outperforms state-of-the-art methods in ML-ZSL.

Key words: Chain-of-thought; Multi-label zero-shot learning; Multimodal reasoning; Large language model
https://doi.org/10.1631/FITEE.2500429 **CLC number:** TP18

1 Introduction

Real-world machine learning applications such as image annotation, music categorization, and medical diagnosis require assigning more than one class label to each input instance. For example, in image annotation, a model may need to assign several labels (e.g., sky, sea, and ship) to a single image. This fundamentally differs from conventional multiclass classification, which assumes that each instance is associated with exactly one label. Developing effective multi-label classification models typically involves additional challenges. In particular, it is necessary not only to accurately associate input instances with multiple relevant labels but also to model and leverage label correlations, which frequently arise due to

‡ Corresponding author
* Project supported by the Key R&D Program of Zhejiang Province (No. 2024C01021), the National Regional Innovation and Development Joint Fund of China (No. U24A20254), and the Leading Talents of Technological Innovation Program of Zhejiang Province (No. 2023R5214)
# Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.2500429) contains supplementary materials, which are available to authorized users
ORCID: Gang CHEN, https://orcid.org/0000-0002-7483-0045
© Zhejiang University Press 2025

the co-occurrence of certain labels in real-world data.

In general, binary relevance (Tsoumakas and Katakis, 2008) is the simplest solution to multi-label classification problems in which the original task is converted into multiple disjoint binary classification problems. However, it cannot model label co-occurrences and may not be preferable. Approaches such as that proposed by Read et al. (2011) take cross-label correlation by assuming label priors, whereas label embedding-based methods (Balasubramanian and Lebanon, 2012; Chen YN and Lin, 2012; Tai and Lin, 2012; Changpinyo et al., 2016, 2017) project both input images and their labels onto a latent space to exploit label correlation. Deep neural network-based methods have also been proposed. BP-MLL (Zhang ML and Zhou, 2006) was the first to propose a loss function for modeling the dependency across labels. Other works have proposed different loss functions (Nam et al., 2014) or architectures (Wei et al., 2014; Wang J et al., 2016; Yeh et al., 2017) to further improve performance.

Multi-label zero-shot learning (ML-ZSL) extends conventional multiclass classification paradigms by addressing a more challenging setting within the zero-shot learning (ZSL) framework. Unlike standard classification tasks, ML-ZSL requires models to simultaneously predict multiple semantic labels for each instance, including novel classes that are absent during training. This poses a complex inference problem that requires the design of advanced transfer mechanisms capable of exploiting semantic relationships between seen and unseen label spaces, as illustrated in Fig. 1. The core innovation of ML-ZSL lies in its ability to generalize learned representations across disjoint label distributions while capturing label co-occurrence patterns, an

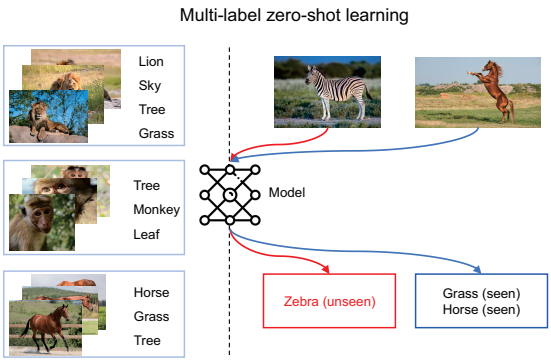

Fig. 1 Examples for multi-label zero-shot learning

essential feature for effective multi-label prediction under zero-shot conditions. Conventional multi-label methods, such as binary relevance or label prior-based approaches, are inherently inadequate for ML-ZSL, because they cannot generalize beyond the set of observed classes.

In contrast, approaches that use label representations in the semantic space, such as label embedding methods, can be easily adapted to ML-ZSL, given label representations of the unseen classes. Label representations are generally obtained from human-annotated attribute vectors that describe the labels of interest either in a specific domain or via distributed word embeddings learned from linguistic resources. Nevertheless, although many ML-ZSL methods have been proposed, including those proposed by Mensink et al. (2014), Fu et al. (2015), Ren et al. (2015), and Zhang Y et al. (2016), existing approaches typically do not take advantage of structured knowledge and reasoning. Humans recognize objects not only by appearance but also by using knowledge of the world learned through experience.

However, some studies on multi-label problems have incorporated structured knowledge. Deng et al. (2014) introduced a graph representation that enforces certain relations between label concepts. Hu et al. (2016) employed recurrent neural networks (RNNs) (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997) to model positive and negative correlations between different concept layers. More recently, Marino et al. (2016) extended neural networks for graphs to efficiently learn a model that reasons about different types of relationships between class labels by propagating information in a knowledge graph (KG). Lee et al. (2018) leveraged structured KGs to model the relationships between labels, thereby enabling the prediction of multiple unseen class labels for each input instance by learning an information propagation mechanism in the semantic label space.

Recently, CLIP-Decoder (Ali and Khan, 2023) and MKT (He et al., 2023) have been proposed. These are novel methods for multi-label zero-shot classification that leverage multimodal representation learning by aligning image and text features using contrastive language–image pretraining (CLIP)-aligned embeddings and vision–language pretraining (VLP) models. They use knowledge distillation and prompt tuning to enhance image–text matching

ability and a two-stream module to capture local and global features, achieving state-of-the-art (SOTA) results in ZSL and generalized ZSL (GZSL) tasks on benchmark datasets. However, these two methods do not fully leverage the information of KGs between labels and CLIP modalities through independent encoders, which limits their ability to capture complex local interactions between modalities.

To address these issues, we propose a graph-augmented multimodal chain-of-thought (GMCoT) reasoning approach. The proposed approach leverages the strengths of native multimodal language models and integrates them with graph-based structures, thereby significantly improving the reasoning process for multi-label prediction. The primary contributions of this work are as follows:

1. We propose the GMCoT reasoning framework that emulates human-like step-by-step inference to facilitate multi-label prediction. By explicitly modeling intermediate reasoning steps, the proposed framework offers a more effective and interpretable solution for complex multi-label classification tasks.

2. To further strengthen the reasoning process, we incorporate label graphs to explicitly model the complex semantic dependencies among labels. This graph-based integration significantly improves the accuracy and coherence of multi-label predictions, especially in critical label co-occurrence scenarios.

3. We conduct extensive experiments on standard benchmark datasets, and the results demonstrate that the proposed GMCoT framework consistently outperforms SOTA methods in ML-ZSL. These findings highlight the effectiveness and generalizability of GMCoT in addressing the unique challenges of multi-label classification tasks.

2 Background and related works

2.1 Multi-label zero-shot learning

The goal of a standard multi-label classification task is to predict a set of labels in an image. A vanilla approach is to train a binary classifier for each label in the training dataset without considering label dependence (Tsoumakas and Katakis, 2007; Read et al., 2011). To capture the label correlation, structure learning (Gong et al., 2014; Wang J et al., 2016) and graph methods (Li Q et al., 2016) are introduced in this task. Recently, vision Transformer (ViT)-based methods have received significant attention because of their ability to capture global dependency (Lanchantin et al., 2021; Liu et al., 2021; Cheng et al., 2021). Although these methods have achieved promising results in multi-label classification, they cannot handle unseen labels, thereby limiting their practical applications.

To identify unseen labels, ZSL typically uses semantic information such as attributes or word embeddings (Mikolov et al., 2013; Xian et al., 2017). In particular, Lampert et al. (2009) proposed two attribute-based paradigms with direct attribute prediction (DAP) and indirect attribute prediction (IAP). The former aims to learn multiple attribute classifiers (Lampert et al., 2014), whereas the latter uses seen class proportions for prediction (Zhang ZM and Saligrama, 2015). Although they can recognize a single unseen label, they cannot handle the multi-label problem. As an extension of ZSL, ML-ZSL is developed to identify multiple seen and unseen labels in an image. The success of this task is based on two key factors: alignment of image embeddings with their relevant label embeddings and the relation between seen and unseen label embeddings. In this regard, Zhang Y et al. (2016) and Ben-Cohen et al. (2021) aimed to find the principal directions of an image along which the relevant labels rank higher. Huynh and Elhamifar (2020) and Narayan et al. (2021) introduced an attention module to capture both local and global features for improved recognition of multiple objects. Gupta et al. (2023) introduced generative adversarial networks to address the problem of multi-label feature synthesis from corresponding multi-label class embeddings.

However, most existing ML-ZSL approaches rely solely on single-modal knowledge, typically extracted from large language models (LLMs). Although these methods are effective in capturing semantic relationships among textual labels, they suffer from the absence of visual grounding, making modeling visual consistency across labels difficult. This limitation significantly hinders their ability to generalize to unseen visual categories. In contrast, the proposed approach explores multimodal knowledge by leveraging VLP models, which jointly encode visual and textual modalities. By aligning image features with corresponding label embeddings in a shared semantic space, our method effectively

captures cross-modal consistency. This enables robust generalization to multiple unseen labels expressed in diverse wordings or textual descriptions, including those with subtle semantic variations.

2.2 Multimodal CoT reasoning with LLMs

The prevalence of image data and associated tasks has driven the extensive application of MCoT, mostly in visual question answering (VQA). Early implementations, such as the interactive prompting visual reasoner (Chen ZF et al., 2023) and Multimodal-CoT (Zhang ZS et al., 2023), established the foundational MCoT framework by generating intermediate rationales before final predictions. Subsequent advancements have further refined this paradigm. For example, the MCoT method proposed by Tan et al. (2024) integrates self-consistency (Wang XZ et al., 2022) with MCoT reasoning, employing word-level majority voting during training to improve the quality of generated rationales. SoT (Aytes et al., 2025) leverages a router model to dynamically select reasoning paradigms (e.g., conceptual chaining, chunked symbolism, and expert lexicons) inspired by human cognitive strategies to enhance reasoning efficiency. CoCoT (Zhang DA et al., 2024) improves multi-image comprehension in multimodal LLMs (MLLMs) through similarity and difference analysis across inputs. Relational MM (Xie et al., 2025) explicitly addresses object relationship modeling via task decomposition. HoT (Yao et al., 2023) extends the graph-of-thought framework by introducing hyperedges to connect multiple reasoning nodes, thereby enhancing multimodal reasoning capabilities.

However, despite these advancements, several key challenges remain in effectively extending MCoT to broader multimodal scenarios such as ML-ZSL. First, most existing MCoT frameworks are tailored for single-question single-answer formats (e.g., VQA), where reasoning is linear and goal-oriented. In contrast, multi-label prediction requires simultaneous reasoning over multiple, potentially correlated labels, requiring a more flexible and parallelizable reasoning structure. Second, many MCoT methods depend heavily on annotated rationales or predefined visual regions of interest, limiting their scalability and adaptability to unseen concepts in zero-shot settings.

3 GMCoT

This section introduces the GMCoT reasoning framework, as illustrated in Fig. 2. The proposed framework draws inspiration from human cognitive processes by decomposing multi-label prediction into a sequence of intermediate reasoning steps. This structured inference paradigm enables the model to progressively refine its predictions in a step-by-step manner, thereby enhancing the effectiveness and interpretability of the model in complex classification tasks. To enrich the reasoning capabilities of GMCoT, we integrate a label graph that explicitly captures the semantic dependencies and

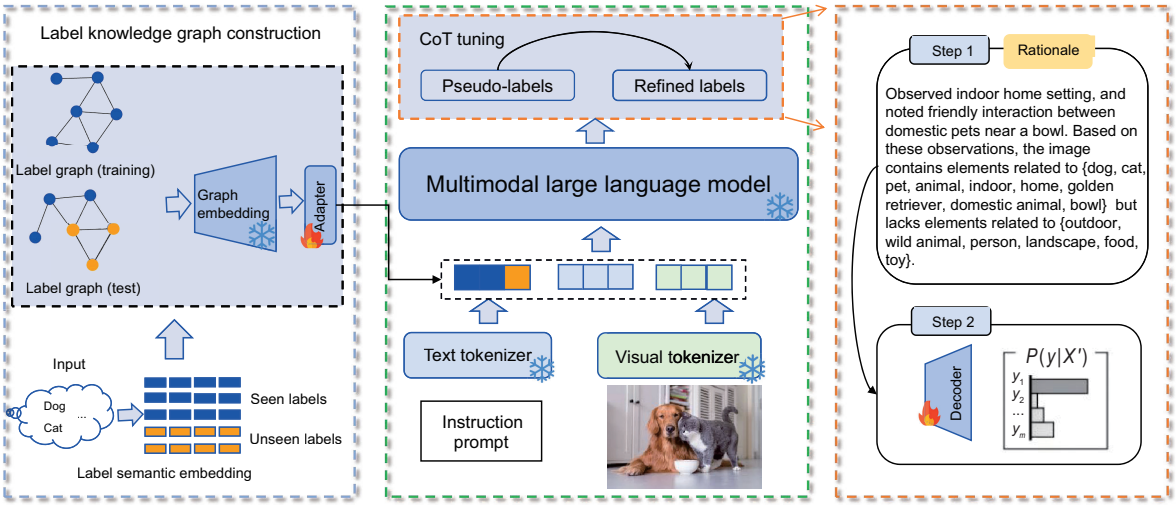

Fig. 2 Graph-augmented multimodal chain-of-thought framework for multi-label zero-shot learning

co-occurrence patterns among labels. This graph-based structure allows the model to reason not only over the visual–textual alignment of each label but also over their relational consistency. By jointly leveraging multimodal signals and structured label relationships, GMCoT achieves more coherent, accurate, and context-aware predictions, particularly in scenarios with high label ambiguity or strong inter-label dependencies, such as zero-shot and fine-grained classification settings.

3.1 Problem formulation

Conventional multi-label classification methods typically assume that all tags to be predicted at test time have already been encountered in training (Wang J et al., 2016; Li Y et al., 2017). However, in realistic annotation scenarios, many tags that were never seen during training are typically encountered at test time; such tags are typically referred to as unseen classes. Addressing the challenge of correctly predicting labels that do not appear in the training set leads to the formulation of the ML-ZSL problem (Mensink et al., 2014; Lee et al., 2018). We formally define the ML-ZSL problem as follows. We consider two disjoint subsets of labels, namely seen and unseen tags. Let S represent the set of seen tags (which appear during training), and U be the set of unseen tags (which do not appear in the training annotations but are desired during testing). Thus, the entire set of labels C can be represented as follows:

$$C = S \cup U,\ S \cap U = \varnothing. \tag{1}$$

Let us denote our labeled training dataset by $\{(I_n, Y_n) \mid n = 1, 2, \ldots, N\}$, where each training example is represented by an image I_n and its corresponding subset of seen tags $Y_n \subseteq S$. Given that we only see labels from S during training, predicting labels from the unseen set U at test time poses a fundamental difficulty. This difficulty arises primarily from the lack of direct supervised information about unseen tags, requiring the approach to leverage auxiliary semantic information (e.g., using semantic embeddings such as textual word embeddings or attribute vectors). To address this challenge, ML-ZSL methods generally exploit semantic representations shared between seen and unseen tags, thereby enabling knowledge transfer from seen to unseen classes (Akata et al., 2016). We assume that each tag $c \in C$ is represented by its corresponding semantic vector embedding:

$$\{v_c\}_{c \in C},\ v_c \in \mathbb{R}^d, \tag{2}$$

where d represents the dimension of semantic embeddings. These embeddings are sourced from auxiliary semantic resources, which increasingly leverage powerful multimodal models such as the text encoder of CLIP. This approach provides a structured semantic representation for tags that is inherently aligned with the visual domain, allowing the model to establish robust relations between seen and unseen classes. Given this problem formulation, the primary challenge in ML-ZSL is how to effectively exploit these semantic representations to generalize knowledge from seen to unseen tags, thereby predicting meaningful multi-label assignments for unseen examples.

3.2 Label knowledge graph construction

Explicitly capturing semantic relationships among labels is essential for effective ML-ZSL, enabling knowledge transfer from seen labels (present during training) to unseen labels (emerging at testing). To this end, we propose explicitly constructed label KGs that encode structured semantic connections between labels.

Formally, we define two separate stage-dependent graphs. During the training stage, we use the graph $G_{\text{train}} = (V_S, \mathcal{E}_S)$, which exclusively comprises nodes corresponding to seen labels V_S. At inference, we construct the graph $G_{\text{test}} = (V_{S\cup U}, \mathcal{E}_{S\cup U})$, which incorporates seen and unseen labels, enabling semantic inference on previously unseen labels.

In ML-ZSL, the label set C comprises a set of seen labels S available during training and unseen labels U appearing only during testing, with $C = S \cup U$ and $S \cap U = \varnothing$.

Each label is initially represented by a semantic embedding obtained from the CLIP model, resulting in a d-dimensional semantic embedding space. Consequently, each label $c \in C$ initially corresponds to an embedding $h_c^{(0)} \in \mathbb{R}^d$, thereby capturing preliminary label semantics.

Each label corresponds to a unique node within the label KG. Consequently, the initial node embedding matrix for all labels is expressed as $H^{(0)} \in \mathbb{R}^{n\times d}$, where $n = |C|$ denotes the total number of labels. The two separate label graphs tailored to the

training and testing stages are formally defined as follows:

Training-stage graph: $G_{\text{train}} = (V_S, \mathcal{E}_S)$ with seen nodes and their respective semantic edges.

Testing-stage graph: $G_{\text{test}} = (V_{S\cup U}, \mathcal{E}_{S\cup U})$, integrating seen and unseen label nodes to facilitate inference.

Initially constructed label graphs typically exhibit sparsity, which potentially hinders information propagation and limits zero-shot generalization. To address this limitation, we augment semantic relations based on embedding similarities. We leverage pairwise cosine similarity between label embeddings to insert additional edges, defined as follows:

$$\begin{aligned} E_{\text{aug}} =& E \cup \{(v_i, v_j) \mid \cos(h_i^{(0)}, h_j^{(0)}) > \epsilon, \\ & \forall v_i, v_j \in V, i \neq j\}, \end{aligned} \tag{3}$$

where ϵ denotes a threshold parameter empirically set at 0.7, balancing graph connectivity and semantic relevance based on validation results.

3.2.1 Semantic graph embedding via pre-training graph convolutional networks

We leverage graph convolutional networks (GCNs) to learn structured semantic embeddings for labels based on the constructed label KG. The obtained label embeddings capture both semantic meaning and structural relationships among labels to facilitate effective ML-ZSL.

To generate rich and robust semantic label embeddings, we introduce a pretraining phase using a dedicated graph neural network (GNN). Specifically, given the entire label set $Y = \{y_1, y_2, \ldots, y_K\}$, we define an initial semantic graph $G = (V, E)$, where nodes V correspond to the labels and edges E denote their semantic relationships, constructed as described in earlier sections.

We then employ a GNN $f_{\text{GNN}}(\cdot; \Theta_{\text{GNN}})$ parameterized by Θ_{GNN} to encode each label node $v_i \in V$ into a fixed-dimensional embedding space:

$$e_i = f_{\text{GNN}}(v_i; \Theta_{\text{GNN}}),\ e_i \in \mathbb{R}^d, \tag{4}$$

where d denotes the embedding dimension. During this pretraining stage, GNN parameters Θ_{GNN} are optimized using a suitable self-supervised or semi-supervised graph-based loss (e.g., node attribute reconstruction, graph embedding alignment, or node classification) to ensure that the embeddings capture meaningful semantic and structural information.

After pretraining, the entire set of learned parameters Θ_{GNN} is frozen to maintain stable semantic representations. The resulting embedding set is defined as follows:

$$E = \{e_1, e_2, \ldots, e_K\},\ e_i \in \mathbb{R}^d,\ i = 1, 2, \ldots, K. \tag{5}$$

During inference, node embeddings for both seen and unseen labels are directly derived from the frozen GNN encoder parameters obtained from pre-training. This freeze-and-reuse approach guarantees semantic consistency and generalizability of label representations, which are critical for handling unseen labels in zero-shot scenarios.

3.2.2 Adapter-based integration with vision-language models

To effectively incorporate the graph-structured semantic embeddings into a multimodal vision–language framework, we propose to use a lightweight adapter module. The adapter module typically comprises a bottleneck architecture and a residual connection, and it is formally defined as follows:

$$\text{Adapter}(x) = x + f_{\text{up}}\left(\sigma(f_{\text{down}}(x))\right), \tag{6}$$

where f_{down} and f_{up} denote the parameterized projection layers, and $\sigma(\cdot)$ denotes a nonlinear activation function (e.g., GELU). This compact design introduces a minimum number of additional parameters, enabling parameter-efficient adapter-based fine-tuning.

This integration strategy affords several key benefits: (1) high parameter efficiency, because adapter modules significantly reduce the number of fine-tuning parameters compared with full model updates; (2) preservation of pretrained knowledge, because freezing the entire pretrained vision–language backbone retains generic prior knowledge; (3) multimodal alignment, because adapter modules effectively bridge modality gaps between graph-based semantic and vision–language embeddings.

During training, only the adapter parameters and specific task-heads are optimized, whereas the pretrained vision–language model parameters are frozen. This enables efficient adaptation to zero-shot scenarios and enhances computational efficiency.

3.3 Multimodal chain-of-thought reasoning

To effectively integrate semantic label embeddings derived from the KG with visual and textual embeddings derived from modality-specific encoders, an MCoT framework is proposed for zero-shot multi-label classification. The methodology for constructing and optimizing this MCoT process is decomposed into two primary operational stages: (1) rationale generation and (2) answer inference, followed by training procedures and zero-shot deployment mechanisms.

3.3.1 Operational workflow

The MCoT framework operates through a two-stage sequential process, as shown in Fig. 3. In the rationale generation stage, the model receives multimodal input $X = \{X^1_{\text{language}}, X_{\text{vision}}, X_{\text{graph}}\}$, where X^1_{language} denotes the initial textual query (e.g., "predict whether the image contains each of the following labels: $\{y_1, y_2, ..., y_K\}$"), X_{vision} denotes the visual input, and X_{graph} denotes the embedding of the label KG. The goal is to learn a rationale generation function $R = F(X)$, where R encodes intermediate reasoning steps.

In the answer inference stage, the generated rationale R is appended to the original language input, forming $X^2_{\text{language}} = X^1_{\text{language}} \circ R$, where $\circ$ denotes concatenation. The updated multimodal input $X' = \{X^2_{\text{language}}, X_{\text{vision}}, X_{\text{graph}}\}$ is then processed by the answer inference component to produce the final multi-label predictions $P = (y|X')$.

3.3.2 Multimodal alignment and interactions

An image–text encoder backbone is instantiated from a large-scale, pretrained multimodal foundation model, e.g., the Mono-InternVL architecture (Luo et al., 2025). The backbone captures modality-specific representations for vision–language inputs. Given an input image–text pair (I, T) and the graph embeddings of the label KG, the pretrained multimodal backbone outputs the following latent representations:

$$z_I = h_{\text{img}}(I; \Theta_{\text{frozen}}) \in \mathbb{R}^{d_{\text{m}}}, \tag{7}$$

$$z_T = h_{\text{txt}}(T; \Theta_{\text{frozen}}) \in \mathbb{R}^{d_{\text{m}}}, \tag{8}$$

$$z_{\text{G}} = h_{\text{emb}}(T; \Theta_{\text{frozen}}) \in \mathbb{R}^{d_{\text{m}}}, \tag{9}$$

where Θ_{frozen} indicates the frozen parameters from pre-training, and d_{m} denotes the multimodal embedding dimension. To facilitate enhanced alignment and interaction between the multimodal and semantic embeddings, lightweight, learnable adapter modules are integrated within each Transformer layer of the pretrained multimodal backbone.

For the lth Transformer layer, the embedding interaction via the adapter module is modeled as follows:

$$z^{(l+1)} = z^{(l)} + f_{\text{adapter}}\left(z^{(l)}; \theta^{(l)}_{\text{adapter}}\right), \tag{10}$$

where $f_{\text{adapter}}(\cdot)$ is typically implemented as a lightweight multilayer perceptron (MLP) or a residual bottleneck structure, and $\theta^{(l)}_{\text{adapter}}$ denotes the trainable parameters of the adapter at layer l. These

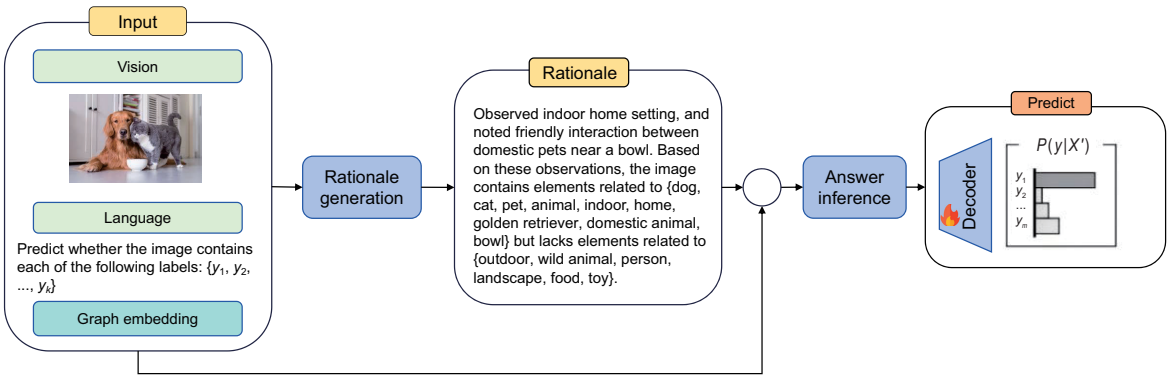

Fig. 3 Overview of the proposed graph-augmented multimodal chain-of-thought framework. The proposed framework involves two stages: (1) a rationale generation stage that combines vision, language, and graph embedding inputs to produce detailed observations about visual content; (2) an answer inference stage that uses the generated rationale and the original inputs to calculate $P(y|x, R)$ and make final predictions. This approach mimics human reasoning by explicitly analyzing visual information before concluding, thereby enhancing both interpretability and accuracy in multimodal tasks

modules act as task-specific transformation layers that introduce a minimum number of additional parameters while preserving the knowledge of the frozen backbone. By injecting learnable modulation within intermediate representations, the adapters enable selective refinement of pretrained multimodal embeddings, thereby facilitating effective knowledge transfer to downstream tasks. This design maintains the intrinsic representational capacity of the base model, ensuring semantic consistency across modalities and enhancing multimodal feature alignment through localized adaptation.

3.3.3 Training via self-distillation and pseudo-label optimization

The proposed framework is trained and optimized through a sophisticated two-stage process that relies on pseudo-labeling and iterative self-distillation, rather than direct supervision on ground-truth labels. This approach is designed to effectively train the two CoT components: the rationale generator and the answer inferer. The entire process is structured as an initial pseudo-label generation phase, followed by an iterative refinement phase.

1. Initial pseudo-label generation

In the first phase, a set of high-quality initial pseudo-labels is generated. In this context, a pseudo-label is defined as a pair $(R_{\text{pseudo}}, A_{\text{pseudo}})$, consisting of a textual rationale and its corresponding set of inferred labels. These are produced using a powerful, pre-existing MLMM, hereafter referred to as the "teacher" model. For each training sample X_n, the teacher model is prompted with a CoT instruction to elicit a detailed reasoning process (the rationale) that culminates in a final label prediction. This process leverages the advanced reasoning capabilities of the teacher model to create a rich supervisory signal that guides the training of our comparatively smaller "student" model. This initial set of pseudo-labels, denoted as $\mathcal{D}^{(0)}_{\text{pseudo}} = \{(X_n, R^{(0)}_{\text{pseudo},n}, A^{(0)}_{\text{pseudo},n})\}_{n=1}^{N}$, serves as the foundational training data.

2. Iterative self-distillation for optimization

Following the initial generation, an iterative self-distillation procedure is employed to progressively refine the pseudo-labels and improve the performance of the student model. At each iteration i, the student model, comprising the rationale generation and answer inference components, is trained using the pseudo-label dataset from the previous iteration, $\mathcal{D}^{(i-1)}_{\text{pseudo}}$.

The training within an iteration is aligned with the two-stage CoT architecture, sequentially addressing each component.

3. Rationale generation training

The rationale generation model is trained to minimize a sequence-to-sequence loss (e.g., cross-entropy) between its generated rationale R_n and the pseudo-rationale $R^{(i-1)}_{\text{pseudo},n}$. This objective is formulated to mimic the reasoning process provided by the teacher model:

$$\mathcal{L}^{(i)}_{\text{rationale}} = -\frac{1}{N}\sum_{n=1}^{N} \log P(R^{(i-1)}_{\text{pseudo},n}|X_n). \tag{11}$$

4. Answer inference training

The answer inference model is trained to predict the final labels. The model processes the input X_n augmented with the rationale R_n generated by the rationale generator of the student model. The supervisory signal is the pseudo-answer $A^{(i-1)}_{\text{pseudo},n}$.

The training objective is a binary cross-entropy (BCE) loss. First, the loss for a single data instance n, denoted as L_n, is calculated by summing the BCE terms over all possible labels in the set $\mathcal{Y}$.

$$\begin{aligned} L_n = -\sum_{y_k\in\mathcal{Y}} \Bigg[& 1[y_k \in A^{(i-1)}_{\text{pseudo},n}] \log p(y_k|X_n, R_n) \\ & + (1 - 1[y_k \in A^{(i-1)}_{\text{pseudo},n}]) \\ & \cdot \log\big(1 - p(y_k|X_n, R_n)\big)\Bigg]. \end{aligned} \tag{12}$$

The total inference loss for the batch, $\mathcal{L}^{(i)}_{\text{inference}}$, is then formulated as the average of these individual instance losses over all N samples:

$$\mathcal{L}^{(i)}_{\text{inference}} = \frac{1}{N}\sum_{n=1}^{N} L_n. \tag{13}$$

Upon completion of the training for iteration i, the optimized student model is then used as the new teacher to generate an updated set of pseudo-labels, $\mathcal{D}^{(i)}_{\text{pseudo}}$. This is accomplished by performing inference on the entire training set using the student model from iteration i. The newly generated rationales and answers are typically more consistent with the representational space of the student model,

serving as a more refined target for the next training iteration, $i+1$. This iterative process continues for a predefined number of iterations or until the pseudo-labels stabilize, allowing the model to distill its knowledge and improve its reasoning and prediction capabilities iteratively. Throughout this process, only the lightweight adapters and the heads of the rationale and inference modules are optimized, while the pretrained backbone parameters are frozen.

3.3.4 Zero-shot inference

During the inference phase, the MCoT framework sequentially executes both operational stages. First, the trained rationale generation model processes the test input to produce structured reasoning steps. Then, these rationales are incorporated into the input for the answer inference model, which leverages semantic embeddings of labels (including those unseen during training) to perform zero-shot classification. Given the multimodal representation of a test input, z_{test}, predictive probabilities for both seen and unseen labels are computed by leveraging their respective semantic embeddings e_j.

$$p(y_j|I_{\text{test}}, T_{\text{test}}) = \sigma\left(\frac{z_{\text{test}}^{\text{T}} e_j}{\tau}\right), \tag{14}$$

where $y_j \in \mathcal{Y}$, and the total label space $\mathcal{Y}$ is defined as the union of the seen label set Y_{S} and the unseen label set Y_{U} (i.e., $\mathcal{Y} = Y_{\text{S}} \cup Y_{\text{U}}$). This inference mechanism facilitates the effective transfer of knowledge from seen to previously unseen labels, mediated by well-structured semantic representations and learned multimodal interactions.

By implementing this two-stage reasoning paradigm, trained via an iterative self-distillation workflow, the proposed MCoT framework systematically leverages pretrained semantic, visual, and textual representations, in conjunction with adapter-based interactions and a targeted optimization strategy (as detailed in Algorithm 1), to achieve robust and accurate zero-shot multimodal predictions.

4 Experiments

4.1 Experimental setup

4.1.1 Datasets

The proposed GMCoT framework is evaluated on two large-scale multi-label benchmark datasets that are widely used in the literature. The NUS-WIDE dataset (Chua et al., 2009) contains 269 648 images annotated with 81 human-verified concept labels and 925 user-generated tags. Following established protocols in ML-ZSL (Huynh and Elhamifar, 2020), we designate the 925 user tags as seen labels for training and the 81 verified concept labels as unseen categories for evaluation. The dataset is partitioned into 161 789 training images and 107 859 testing images, maintaining the original distribution characteristics.

Algorithm 1 Iterative self-distillation for multimodal chain-of-thought

Require: training data $\mathcal{D}_{\text{train}} = \{X_n\}_{n=1}^{N}$; test data $\mathcal{D}_{\text{test}}$; initial teacher model M_{teacher}; number of self-distillation iterations M.
Ensure: zero-shot predictions Y^* for $\mathcal{D}_{\text{test}}$.

// **Phase 1: initialization**
Construct label KG G and compute graph embeddings H_{graph}.
Initialize student model $M_{\text{student}}(\theta)$ with frozen backbone and trainable adapters.
Generate the initial pseudo-label dataset $\mathcal{D}_{\text{pseudo}}^{(0)} = \{(R_{\text{pseudo},n}^{(0)}, A_{\text{pseudo},n}^{(0)})\}_{n=1}^{N}$.
for each sample X_n in $\mathcal{D}_{\text{train}}$ **do**
  $(R_{\text{pseudo},n}^{(0)}, A_{\text{pseudo},n}^{(0)}) \leftarrow M_{\text{teacher}}(X_n)$ /* generate via CoT prompting */
end for

// **Phase 2: iterative self-distillation training**
for $i = 0$ to $M-1$ **do**
  /* Train the student model using the pseudo-labels from the previous iteration */
  Train the rationale generator of M_{student} on target rationales $\{R_{\text{pseudo},n}^{(i)}\}$.
  Train the answer inferer of M_{student} on target answers $\{A_{\text{pseudo},n}^{(i)}\}$.
  Update trainable parameters θ of M_{student}.
  /* Generate a new, refined set of pseudo-labels for the next iteration */
  Let M_{student} become the new teacher.
  Initialize empty dataset $\mathcal{D}_{\text{pseudo}}^{(i+1)}$.
  for each sample X_n in $\mathcal{D}_{\text{train}}$ **do**
    $(R_{\text{pseudo},n}^{(i+1)}, A_{\text{pseudo},n}^{(i+1)}) \leftarrow M_{\text{student}}(X_n)$.
    Add $(R_{\text{pseudo},n}^{(i+1)}, A_{\text{pseudo},n}^{(i+1)})$ to $\mathcal{D}_{\text{pseudo}}^{(i+1)}$.
  end for
end for

// **Phase 3: zero-shot inference**
Let $M_{\text{final}} \leftarrow M_{\text{student}}$ be the fully trained model.
for each test sample X_{test} in $\mathcal{D}_{\text{test}}$ **do**
  Generate rationale: $R_{\text{test}} \leftarrow M_{\text{final}}(X_{\text{test}})$.
  Infer the final answer: $A_{\text{test}} \leftarrow M_{\text{final}}(X_{\text{test}}, R_{\text{test}})$.
  Store the prediction.
end for
return final predictions Y^*.

For a substantially more challenging benchmark, we use the Open Images dataset (V4) (Kuznetsova et al., 2020), which contains approximately 9 million training images and 125 456 test images. In accordance with previous research, 7186 labels that appear in more than 100 training samples are designated as seen categories. For evaluation, the top 400 most frequently occurring test labels that are absent from the training set are selected as unseen categories, creating a realistic and challenging zero-shot scenario.

4.1.2 Evaluation protocols and metrics

Experiments are conducted under two evaluation protocols: (1) standard ML-ZSL, where models are trained on seen labels and evaluated exclusively on unseen labels, and (2) the more challenging generalized ML-ZSL task, where models must predict both seen and unseen labels during testing despite being trained only on seen labels. For NUS-WIDE, we treat the 81 human-verified concepts as the unseen label set $\mathcal{U}$, whereas the seen label set $\mathcal{S}$ comprises 925 user-generated tags after removing 75 duplicated entries.

To ensure comprehensive and fair comparison with existing approaches, we use two standard evaluation metrics. Mean average precision (mAP) evaluates the label-wise ranking accuracy across the entire test set, providing a comprehensive assessment of the ability of the model to correctly prioritize relevant labels. This metric is particularly valuable in multi-label scenarios because it accounts for both precision and recall across different threshold values. We also report the F1-score at top-K predictions, which measures image-wise label ranking accuracy by comparing the top-K predicted labels with ground-truth annotations. Following established protocols, we report results with $K \in \{3, 5\}$ for NUS-WIDE and $K \in \{10, 20\}$ for Open Images, reflecting the different average numbers of labels per image in these datasets.

4.1.3 Baseline methods

We compare the proposed approach with six SOTA multi-label classification approaches. LESA (Chen ZM et al., 2019) introduces joint class-aware map disentangling and label correlation embedding to enhance multi-label recognition. BiAM (Yang et al., 2023) leverages bidirectional attention for multimodal training, improving label–image alignment in speech–text scenarios. SDL (Ben-Cohen et al., 2021) uses semantic diversity learning to address class imbalance in zero-shot multi-label classification.

More recent approaches include ML-Decoder (Ridnik et al., 2023), which proposes a Transformer-based classification head with scalable group decoding for efficient multi-label prediction, and CLIP-Decoder (Ali and Khan, 2023), which introduces a multimodal framework that aligns CLIP-based visual and textual representations through Transformer decoder layers, combining classification and CLIP alignment losses to improve zero-shot performance. MKT (He et al., 2023) handles open-vocabulary multi-label classification by leveraging multimodal knowledge from VLP models, employing knowledge distillation to align image and label embeddings, and integrating prompt tuning to adapt textual embeddings for classification tasks. Finally, to contextualize our work against the latest generation of large-scale models, we include Qwen-2.5-VL as a powerful prompting-based baseline. As a SOTA MLLM, Qwen-2.5-VL is evaluated in a zero-shot setting. The model is presented with the test image and the complete set of possible labels for the task and is then prompted to output a list of all labels it identifies within the image. This approach assesses the out-of-the-box performance of the model without any task-specific fine-tuning.

4.2 Implementation process

The detailed experimental settings and implementation specifics of the proposed pipeline are systematically described in the supplementary materials (Section 1).

4.2.1 Label graph construction and multimodal integration

To model explicit semantic relations among labels, distinct label graphs for training and evaluation are constructed. To initialize node features with representations that are both semantically rich and inherently aligned with the visual domain, the text encoder from a pretrained CLIP model is employed. This method is selected over conventional static word vectors (e.g., GloVe) because it provides powerful,

context-aware embeddings that are better suited for multimodal tasks. The resulting 768-dimensional embeddings are then further refined and adapted to the specifics of the downstream task through a two-layer GCN (Jiang et al., 2019). The GCN uses hidden layers of size 512 to process the initial embeddings and fine-tune them for the target Mono-InternVL multimodal space. Symmetrically normalized adjacency matrices with self-loops are used to stabilize training and facilitate effective aggregation.

To enhance semantic richness and mitigate sparsity in label graphs, we augment the graph structure by adding edges between labels with cosine similarity above the threshold $\epsilon = 0.7$, which is empirically selected via validation to balance connectivity and relevance. GCN-generated label embeddings are further refined using a two-layer MLP adapter, which maps 768-dimensional inputs to a 1024-dimensional space and projects them back to 768 dimensions. Each layer uses ReLU activation and a dropout rate of 0.2. The adapter is optimized with Mono-InternVL using consistent training configurations to ensure semantic alignment. For multimodal inputs, images are resized and normalized to 224×224 pixels following Mono-InternVL preprocessing, whereas text inputs, comprising prompts and metadata, are tokenized with a 128-token limit using the tokenizer of the model.

4.2.2 Multimodal large language model

Mono-InternVL is adopted as the underlying multimodal large language backbone because of its effectiveness in multimodal reasoning tasks. Mono-InternVL inherently integrates textual and visual modalities using Transformer-based architectures trained on extensive multimodal data. The pre-trained Mono-InternVL backbone used in our experiments employs a ViT-based visual tokenizer and a GPT-like textual tokenizer for embedding visual and textual inputs into a unified latent embedding space, respectively. To achieve effective MCoT reasoning, we fine-tune Mono-InternVL using specially designed instructional prompts and reasoning-driven pseudo-label supervision.

4.2.3 Chain-of-thought fine-tuning

The Mono-InternVL model is fine-tuned using an MCoT strategy through a two-stage pseudo-label refinement process. In the first stage, initial pseudo-labels, each comprising a rationale and its corresponding answer, are generated. This is accomplished via inference using a meticulously designed instructional question answering template. This template acts as a structural scaffold, guiding the model to follow a specific reasoning pathway (e.g., "visual observations," then "logical reasoning," and finally "conclusion"). Although the template dictates the high-level structure of the thought process, the textual content within each section is generated freely by the model, tailored to each specific input. This process yields structured rationales and provides initial soft supervision for subsequent training. In the second stage, a self-distillation mechanism iteratively refines these labels: predictions from the previous iteration serve as soft targets for a new training round, enabling the model to incrementally enhance its reasoning capabilities.

To ensure stability and robustness during CoT fine-tuning, we set the learning rate using a cosine learning rate scheduler initialized to 2×10^{-5}, which gradually decays to 1×10^{-6}. The fine-tuning procedure runs for 15 epochs, leveraging the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay set to 1×10^{-4}. We employ a batch size of 32 across four NVIDIA Tesla V100 GPUs, each with 32 GB of memory. Gradient clipping at a norm value of 1.0 is used to alleviate potential exploding gradients.

4.3 Evaluation results and analysis

The efficacy of the proposed GMCoT framework is systematically evaluated on two multi-label benchmark datasets, NUS-WIDE and Open Images, under both conventional ZSL and GZSL protocols. The experimental results validate that the proposed GMCoT framework consistently outperforms SOTA methods across all evaluation settings. As shown in Table 1, under the conventional ZSL setting on the NUS-WIDE dataset, GMCoT achieves SOTA performance, with F1-score of 35.2% and 32.6% at $K=3$ and 5, respectively. As shown in Table 2, the performance advantage of GMCoT generalizes well to more challenging datasets, particularly the Open Images benchmark, which features a significantly larger and more diverse label space, along with complex visual scenes. Under the ZSL configuration, GMCoT achieves an F1-score of 20.8% at $K = 10$, accompanied by an exceptionally high recall of 88.5%. As shown in Tables 3 and 4, under the more challenging

Table 1 ZSL results obtained on the NUS-WIDE dataset. Performance is measured in terms of precision (P), recall (R), and F1-score at $K=3,5$, and mean average precision (mAP), all in %

Method	P@3	R@3	F1@3	P@5	R@5	F1@5	mAP
LESA (M=10)	25.7	41.1	31.6	19.7	52.5	28.7	19.4
ZS-SDL	24.2	41.3	30.5	18.8	53.4	27.8	25.9
BiAM	26.6	42.5	32.7	20.5	54.6	29.8	25.9
ML-Decoder	28.2	43.2	34.1	22.3	55.1	30.8	31.1
CLIP-Decoder	28.6	43.5	34.8	**22.7**	55.5	31.1	33.4
MKT	27.7	44.3	34.1	21.4	57.0	31.1	37.6
Qwen-2.5-VL (prompting)	28.1	44.8	34.8	22.5	57.2	32.3	38.8
GMCoT	**28.7**	**45.5**	**35.2**	**22.7**	**57.8**	**32.6**	**39.4**

The best results are in bold

Table 2 ZSL results obtained on the Open Images dataset. Performance is measured in terms of precision (P), recall (R), and F1-score at $K=10,20$, and mean average precision (mAP), all in %

Method	P@10	R@10	F1@10	P@20	R@20	F1@20	mAP
LESA (M=10)	0.7	25.6	1.4	0.5	37.4	1.0	41.7
ZS-SDL	6.1	47.0	10.7	4.4	68.1	8.3	62.9
BiAM	3.9	30.7	7.0	2.7	41.9	5.5	65.6
ML-Decoder	9.2	82.8	17.5	6.4	90.7	10.4	64.8
CLIP-Decoder	11.1	85.3	20.1	6.2	93.3	12.1	67.3
MKT	11.1	86.8	19.7	6.1	94.7	11.4	68.1
Qwen-2.5-VL (prompting)	11.2	87.2	20.2	6.8	94.1	12.2	68.9
GMCoT	**11.9**	**88.5**	**20.8**	**7.1**	**95.8**	**12.5**	**69.6**

The best results are in bold

Table 3 GZSL results obtained on the NUS-WIDE dataset. Performance is measured in terms of precision (P), recall (R), and F1-score at $K=3,5$, and mean average precision (mAP), all in %

Method	P@3	R@3	F1@3	P@5	R@5	F1@5	mAP
LESA (M=10)	23.6	10.4	14.4	19.8	14.6	16.8	5.6
ZS-SDL	27.7	13.9	18.5	23.0	19.3	21.0	12.1
BiAM	25.2	11.1	15.4	21.6	15.9	18.2	9.4
ML-Decoder	27.1	17.6	23.3	21.1	23.2	26.1	19.9
CLIP-Decoder	27.4	18.3	24.8	22.2	23.7	27.5	23.8
MKT	35.9	15.8	22.0	29.9	22.0	25.4	18.3
Qwen-2.5-VL (prompting)	36.1	16.5	22.2	30.1	22.9	25.7	24.5
GMCoT	**36.7**	**19.2**	**26.3**	**30.9**	**24.2**	**28.1**	**25.2**

The best results are in bold

Table 4 GZSL results obtained on the Open Images dataset. Performance is measured in terms of precision (P), recall (R), and F1-score at $K=10,20$, and mean average precision (mAP), all in %

Method	P@10	R@10	F1@10	P@20	R@20	F1@20	mAP
LESA (M=10)	16.2	18.9	17.4	10.2	23.9	14.3	45.4
ZS-SDL	35.3	40.8	37.8	23.6	54.5	32.9	75.3
BiAM	13.8	15.9	14.8	9.7	22.3	14.8	81.7
ML-Decoder	35.8	41.7	38.8	23.7	56.4	34.1	75.4
CLIP-Decoder	38.2	44.5	41.1	26.2	**59.3**	36.3	78.2
MKT	37.8	43.6	40.5	25.4	58.5	35.4	81.4
Qwen-2.5-VL (prompting)	38.4	44.6	41.3	26.4	58.8	36.3	82.1
GMCoT	**38.7**	**45.1**	**42.7**	**26.6**	59.2	**36.4**	**82.5**

The best results are in bold

GZSL paradigm, where the model must correctly distinguish between seen and unseen labels during inference, the proposed GMCoT framework consistently maintains its leading performance across both datasets, demonstrating strong generalization without sacrificing discriminative power on seen classes.

The consistent performance of GMCoT is attributed to the integration of three key components: (1) structured KG embeddings to model latent semantic relationships among labels; (2) an adapter-based architecture for effective multimodal fusion; (3) a CoT reasoning mechanism that refines predictions by traversing the label KG. These innovations work synergistically to address fundamental challenges in ZSL, such as visual–semantic misalignment and insufficient modeling of inter-label dependencies. More comprehensive analysis and additional results are provided in the supplementary materials (Section 2).

4.4 Fairness and bias analysis

To assess the fairness of our model across different label distributions, we analyze performance on frequent versus rare labels in the NUS-WIDE dataset (see the supplementary materials, Section 3).

5 Ablation studies

A comprehensive ablation study is conducted to isolate and quantify the contributions of the proposed components. The effects of the graph embedding module and the CoT fine-tuning procedure on the overall model performance are systematically evaluated. The detailed results, including the experimental setup and specific evaluation metrics, are presented in the supplementary materials (Section 4).

6 Conclusions

We propose the GMCoT reasoning framework that integrates MLLMs with structured label graphs, enabling a human-like step-by-step reasoning process. The incorporation of the label KG effectively captures semantic label dependencies, significantly improving multi-label prediction. Extensive experiments on standard ML-ZSL benchmarks demonstrate that GMCoT consistently outperforms SOTA methods, particularly in scenarios requiring fine-grained semantic understanding and complex inter-label reasoning. Future work will explore the integration of advanced GNN architectures to enhance relational modeling and extend the proposed framework to generalized zero-shot and open-world classification settings.

Contributors

Xiang WEN designed the algorithm. Xiang WEN and Haobo WANG drafted the paper. Ke CHEN, Tianlei HU, and Gang CHEN polished, optimized, revised, and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Akata Z, Perronnin F, Harchaoui Z, et al., 2016. Label-embedding for image classification. *IEEE Trans Patt Anal Mach Intell*, 38(7):1425-1438. https://doi.org/10.1109/TPAMI.2015.2487986

Ali M, Khan S, 2023. CLIP-Decoder: ZeroShot multilabel classification using multimodal CLIP aligned representations. Proc IEEE/CVF Int Conf on Computer Vision Workshops, p.4677-4681. https://doi.org/10.1109/ICCVW60793.2023.00505

Aytes SA, Baek J, Hwang SJ, 2025. Sketch-of-thought: efficient LLM reasoning with adaptive cognitive-inspired sketching. https://doi.org/10.48550/arXiv.2503.05179

Balasubramanian K, Lebanon G, 2012. The landmark selection method for multiple output prediction. https://doi.org/10.48550/arXiv.1206.6479

Ben-Cohen A, Zamir N, Ben-Baruch E, et al., 2021. Semantic diversity learning for zero-shot multi-label classification. Proc IEEE/CVF Int Conf on Computer Vision, p.620-630. https://doi.org/10.1109/ICCV48922.2021.00068

Changpinyo S, Chao WL, Gong BQ, et al., 2016. Synthesized classifiers for zero-shot learning. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5327-5336. https://doi.org/10.1109/CVPR.2016.575

Changpinyo S, Chao WL, Sha F, 2017. Predicting visual exemplars of unseen classes for zero-shot learning. Proc IEEE Int Conf on Computer Vision, p.3496-3505. https://doi.org/10.1109/ICCV.2017.376

Chen YN, Lin HT, 2012. Feature-aware label space dimension reduction for multi-label classification. Proc 26th Int Conf on Neural Information Processing Systems, p.1529-1537.

Chen ZF, Zhou QH, Shen YK, et al., 2023. See, think, confirm: interactive prompting between vision and language models for knowledge-based visual reasoning. https://doi.org/10.48550/arXiv.2301.05226

Chen ZM, Wei XS, Wang P, et al., 2019. Multi-label image recognition with graph convolutional networks. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.5172-5181. https://doi.org/10.1109/CVPR.2019.00532

Cheng X, Lin HZ, Wu XY, et al., 2021. MlTr: multi-label classification with Transformer. https://doi.org/10.48550/arXiv.2106.06195

Chua TS, Tang JH, Hong RC, et al., 2009. NUS-WIDE: a real-world web image database from National University of Singapore. Proc ACM Int Conf on Image and Video Retrieval, Article 48. https://doi.org/10.1145/1646396.1646452

Deng J, Ding N, Jia YQ, et al., 2014. Large-scale object classification using label relation graphs. 13th European Conf on Computer Vision, p.48-64. https://doi.org/10.1007/978-3-319-10590-1_4

Fu YW, Yang YX, Hospedales TM, et al., 2015. Transductive multi-class and multi-label zero-shot learning. https://doi.org/10.48550/arXiv.1503.07884

Gong YC, Jia YQ, Leung T, et al., 2014. Deep convolutional ranking for multilabel image annotation. https://doi.org/10.48550/arXiv.1312.4894

Gupta A, Narayan S, Khan S, et al., 2023. Generative multi-label zero-shot learning. https://doi.org/10.48550/arXiv.2101.11606

He SN, Guo TA, Dai T, et al., 2023. Open-vocabulary multi-label classification via multi-modal knowledge transfer. 37th AAAI Conf on Artificial Intelligence, p.808-816. https://doi.org/10.1609/aaai.v37i1.25159

Hochreiter S, Schmidhuber J, 1997. Long short-term memory. *Neur Comput*, 9(8):1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu HX, Zhou GT, Deng ZW, et al., 2016. Learning structured inference neural networks with label relations. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2960-2968. https://doi.org/10.1109/CVPR.2016.323

Huynh D, Elhamifar E, 2020. A shared multi-attention framework for multi-label zero-shot learning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.8773-8783. https://doi.org/10.1109/CVPR42600.2020.00880

Jiang B, Zhang Z, Lin D, et al., 2019. Semi-supervised learning with graph learning-convolutional networks. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11313-11320.

Kuznetsova A, Rom H, Alldrin N, et al., 2020. The Open Images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *Int J Comput Vis*, 128(7):1956-1981. https://doi.org/10.1007/s11263-020-01316-z

Lampert CH, Nickisch H, Harmeling S, 2009. Learning to detect unseen object classes by between-class attribute transfer. IEEE Conf on Computer Vision and Pattern Recognition, p.951-958. https://doi.org/10.1109/CVPR.2009.5206594

Lampert CH, Nickisch H, Harmeling S, 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans Patt Anal Mach Intell*, 36(3):453-465. https://doi.org/10.1109/TPAMI.2013.140

Lanchantin J, Wang TL, Ordonez V, et al., 2021. General multi-label image classification with Transformers. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.16473-16483. https://doi.org/10.1109/CVPR46437.2021.01621

Lee CW, Fang W, Yeh CK, et al., 2018. Multi-label zero-shot learning with structured knowledge graphs. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1576-1585. https://doi.org/10.1109/CVPR.2018.00170

Li Q, Qiao MY, Bian W, et al., 2016. Conditional graphical Lasso for multi-label image classification. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2977-2986. https://doi.org/10.1109/CVPR.2016.325

Li Y, Song Y, Luo J, 2017. Improving multi-label classification with missing labels by learning visual and semantic embeddings. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.5184-5192.

Liu SL, Zhang L, Yang X, et al., 2021. Query2Label: a simple Transformer way to multi-label classification. https://doi.org/10.48550/arXiv.2107.10834

Luo G, Yang X, Dou W, et al., 2025. Mono-InternVL: pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. Proc Computer Vision and Pattern Recognition Conf, p.24960-24971.

Marino K, Salakhutdinov R, Gupta A, 2016. The more you know: using knowledge graphs for image classification. https://doi.org/10.48550/arXiv.1612.04844

Mensink T, Gavves E, Snoek CGM, 2014. COSTA: co-occurrence statistics for zero-shot classification. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2441-2448. https://doi.org/10.1109/CVPR.2014.313

Mikolov T, Sutskever I, Chen K, et al., 2013. Distributed representations of words and phrases and their compositionality. Proc 27th Int Conf on Neural Information Processing Systems, p.3111-3119.

Nam J, Kim J, Loza Mencía E, et al., 2014. Large-scale multi-label text classification—revisiting neural networks. European Conf on Machine Learning and Knowledge Discovery in Databases, p.437-452. https://doi.org/10.1007/978-3-662-44851-9_28

Narayan S, Gupta A, Khan F, et al., 2021. Discriminative region-based multi-label zero-shot learning. Proc IEEE/CVF Int Conf on Computer Vision, p.8711-8720. https://doi.org/10.1109/ICCV48922.2021.00861

Read J, Pfahringer B, Holmes G, et al., 2011. Classifier chains for multi-label classification. *Mach Learn*, 85(3):333-359. https://doi.org/10.1007/s10994-011-5256-5

Ren Z, Jin HL, Lin Z, et al., 2015. Multi-instance visual-semantic embedding. https://doi.org/10.48550/arXiv.1512.06963

Ridnik T, Sharir G, Ben-Cohen A, et al., 2023. ML-Decoder: scalable and versatile classification head. Proc IEEE/CVF Winter Conf on Applications of Computer Vision, p.32-41. https://doi.org/10.1109/WACV56688.2023.00012

Schuster M, Paliwal KK, 1997. Bidirectional recurrent neural networks. *IEEE Trans Signal Process*, 45(11):2673-2681. https://doi.org/10.1109/78.650093

Tai F, Lin HT, 2012. Multilabel classification with principal label space transformation. *Neur Comput*, 24(9):2508-2542. https://doi.org/10.1162/NECO_a_00320

Tan C, Wei JX, Gao ZY, et al., 2024. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. 18th European Conf on Computer Vision, p.305-322. https://doi.org/10.1007/978-3-031-73661-2_17

Tsoumakas G, Katakis I, 2007. Multi-label classification: an overview. *Int J Data Warehous Min*, 3(3):1-13. https://doi.org/10.4018/jdwm.2007070101

Tsoumakas G, Katakis I, 2008. Multi-label classification: an overview. In: Wang J (Ed.), Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications. IGI Global Scientific Publishing, Hershey, PA, USA, p.64-74. https://doi.org/10.4018/978-1-59904-951-9.ch006

Wang J, Yang Y, Mao JH, et al., 2016. CNN-RNN: a unified framework for multi-label image classification. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.2285-2294. https://doi.org/10.1109/CVPR.2016.251

Wang XZ, Wei J, Schuurmans D, et al., 2022. Self-consistency improves chain of thought reasoning in language models. https://doi.org/10.48550/arXiv.2203.11171

Wei YC, Xia W, Huang JS, et al., 2014. CNN: single-label to multi-label. https://doi.org/10.48550/arXiv.1406.5726

Xian YQ, Schiele B, Akata Z, 2017. Zero-shot learning—the good, the bad and the ugly. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.3077-3086. https://doi.org/10.1109/CVPR.2017.328

Xie C, Liang S, Li J, et al., 2025. RelationLMM: large multimodal model as open and versatile visual relationship generalist. *IEEE Trans Patt Anal Mach Intell*, 47(5):3515-3529. https://doi.org/10.1109/TPAMI.2025.3531452

Yang YH, Xu HH, Huang H, et al., 2023. Speech-text based multi-modal training with bidirectional attention for improved speech recognition. IEEE Int Conf on Acoustics, Speech and Signal Processing, p.1-5. https://doi.org/10.1109/ICASSP49357.2023.10096726

Yao FL, Tian CY, Liu JT, et al., 2023. Thinking like an expert: multimodal hypergraph-of-thought (HoT) reasoning to boost foundation modals. https://doi.org/10.48550/arXiv.2308.06207

Yeh CK, Wu WC, Ko WJ, et al., 2017. Learning deep latent space for multi-label classification. Proc 31st AAAI Conf on Artificial Intelligence, p.2838-2844. https://doi.org/10.1609/aaai.v31i1.10769

Zhang DA, Yang JM, Lyu HJ, et al., 2024. CoCoT: contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. https://doi.org/10.48550/arXiv.2401.02582

Zhang ML, Zhou ZH, 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Trans Knowl Data Eng*, 18(10):1338-1351. https://doi.org/10.1109/TKDE.2006.162

Zhang Y, Gong BQ, Shah M, 2016. Fast zero-shot image tagging. IEEE Conf on Computer Vision and Pattern Recognition, p.5985-5994. https://doi.org/10.1109/CVPR.2016.644

Zhang ZM, Saligrama V, 2015. Zero-shot learning via semantic similarity embedding. Proc IEEE Int Conf on Computer Vision, p.4166-4174. https://doi.org/10.1109/ICCV.2015.474

Zhang ZS, Zhang A, Li M, et al., 2023. Multimodal chain-of-thought reasoning in language models. https://doi.org/10.48550/arXiv.2302.00923

List of supplementary materials

1 Implementation process

2 Evaluation results and analysis

3 Fairness and bias analysis

4 Ablation studies

Fig. S1 Performance comparison across different visual attributes on the NUS-WIDE dataset (ZSL setting)

Fig. S2 Performance comparison across label frequency groups on the NUS-WIDE dataset (ZSL setting)

Table S1 Ablation results on the test set