Xiang WEN, Haobo WANG, Ke CHEN, Tianlei HU, Gang CHEN, 2025. GMCoT: a graph-augmented multimodal chain-of-thought reasoning framework for multi-label zero-shot learning. *Frontiers of Information Technology & Electronic Engineering*, 26(12):2623-2637. https://doi.org/10.1631/FITEE.2500429

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

Key words: Chain-of-thought; Multi-label zero-shot learning; Multimodal reasoning; Large language model

Corresponding author: Gang CHEN
E-mail: cg@zju.edu.cn
ORCID: https://orcid.org/0000-0002-7483-0045

Motivation

The Problem: Multi-Label Zero-Shot Learning (ML-ZSL)

- Task: Assigning multiple labels to an image (e.g., "sky", "sea", and "ship") where some labels (unseen classes) never appear during training.
- Difference from standard classification: Requiring the prediction of a set of labels rather than a single class, which can generalize to completely new concepts.

Key Challenges:

- Semantic Gap: Difficulty in effectively transferring knowledge between different modalities (visual features vs. textual semantics).
- Complex Label Correlations: Labels in the real world are highly correlated (e.g., "car" and "road" co-occur). Standard binary relevance methods fail to model these dependencies.

Limitations of Existing Methods:

- Often rely on single-modal knowledge.
- Lack visual grounding for unseen concepts.
- Do not utilize structured reasoning (human-like inference).

Main idea: GMCoT Framework

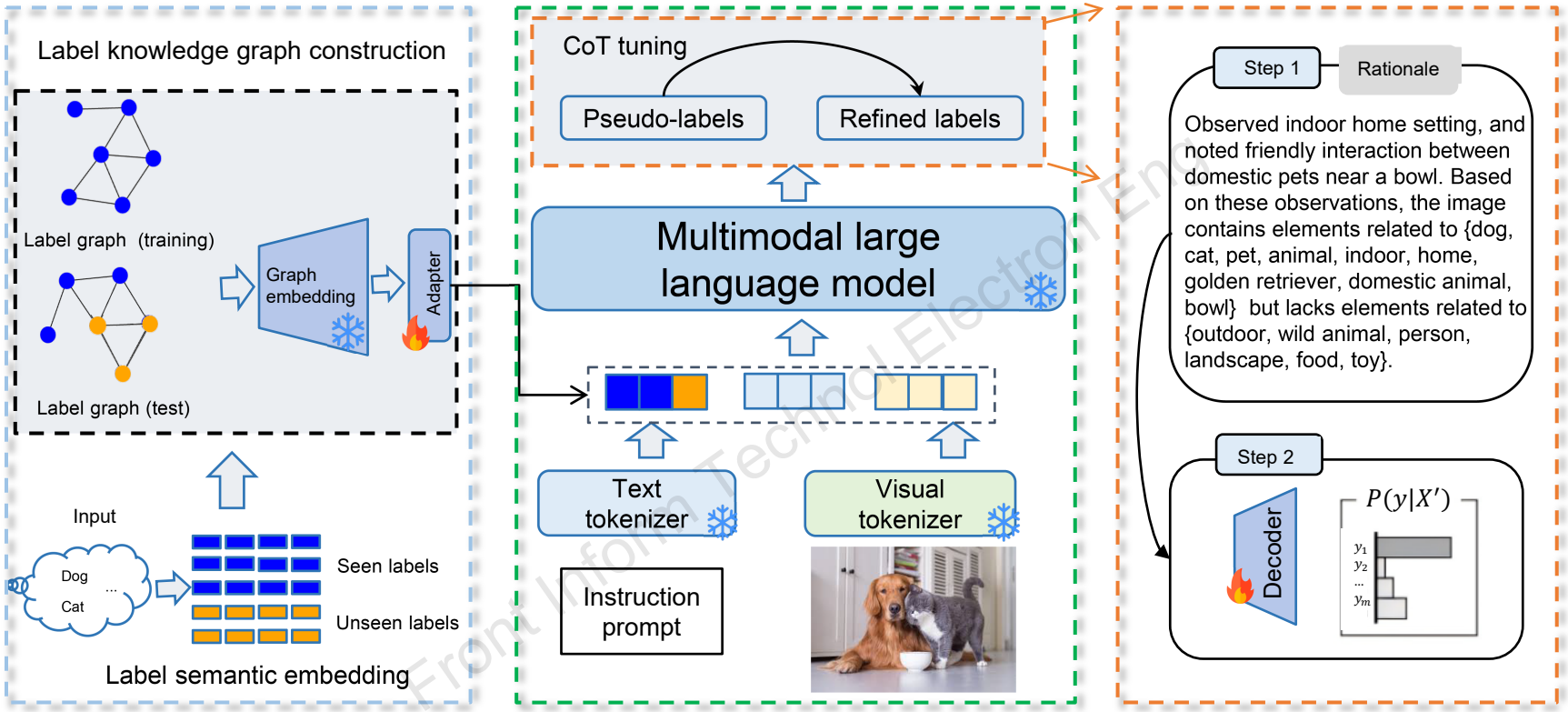
Core Innovation:

- GMCoT (Graph-augmented Multimodal Chain-of-Thought): A novel framework that combines valid multimodal Large Language Models (MLLMs) with graph-based structured knowledge.

Key Approaches:

- Human-like Reasoning: Imitates step-by-step reasoning (Chain-of-Thought) to decompose complex visual recognition into intermediate observations and final conclusions.
- Graph Augmentation: Explicitly integrates a Label Knowledge Graph to capture semantic relationships between seen and unseen labels.
- Synergy: Leverages the generative power of LLMs while constraining and guiding predictions using structured graph embeddings.

Framework



Two-Stage Graph-Augmented Framework

Built on Mono-InternVL with lightweight adapters, our approach fuses graph embeddings in a two-stage pipeline. It first generates natural language rationales to interpret visual content, and then utilizes these rationales to predict final labels, enabling robust zero-shot classification with high parameter efficiency.

Method: Label Knowledge Graph

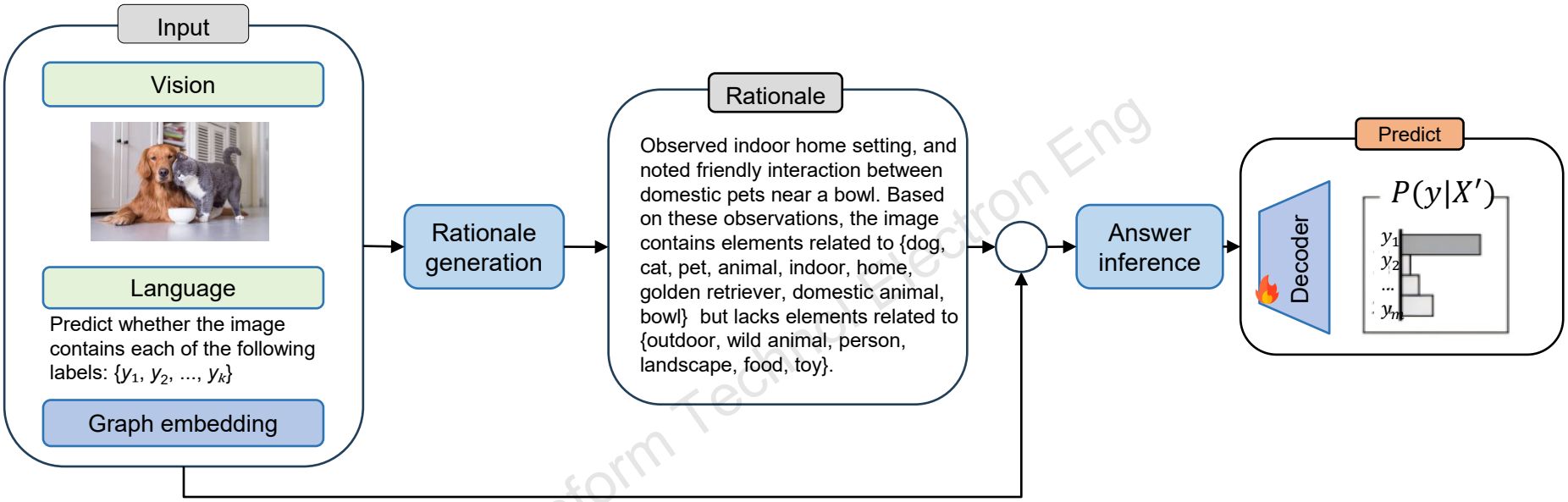
Construction:

- Nodes: Represent all labels (Set $C = S \cup U$), where S represents seen labels and U unseen labels.
- Embeddings: Initialized using the CLIP text encoder to ensure visual-semantic alignment.
- Graph Neural Network (GCN): Refines embeddings to capture structural relationships.

Augmentation Strategy:

- Edges: Created based on semantic relations.
- Enhancement: Additional edges are added based on pairwise cosine similarity ($> \epsilon$) between label embeddings.
- Purpose: Enables effective information propagation from seen classes (training) to unseen classes (inference), bridging the zero-shot gap.

Method: Graph-Augmented CoT



1. Input Fusion: The model integrates visual and textual features with structured graph embeddings to capture semantic correlations between seen and unseen labels.

2. Two-Stage Reasoning: It employs a Chain-of-Thought pipeline that first generates natural language rationales to bridge the semantic gap, guiding more accurate final predictions.

Experimental Setup

Datasets:

1. NUS-WIDE:
   - Training: 925 user-generated tags (Seen).
   - Testing: 81 human-verified concept labels (Unseen).
2. Open Images (V4):
   - A more challenging, large-scale dataset.
   - Training: 7186 labels.
   - Testing: Top 400 frequent labels not present in training.

Evaluation Protocols:

- Zero-Shot Learning (ZSL): Train on seen; evaluate only on unseen.
- Generalized ZSL (GZSL): Train on seen; evaluate on both seen and unseen (more realistic).

Metrics:

- Mean Average Precision (mAP)
- F1-score at Top-*K* (F1@*K*)

Major results

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

Major results

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

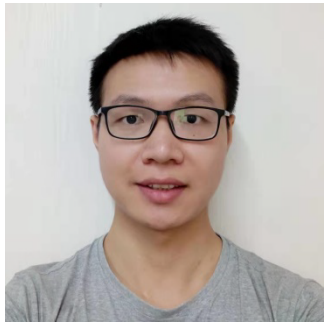
Conclusions

Summary:

- Framework: Proposed GMCoT, a reasoning framework acting as a bridge between visual perception and logical label inference.
- Integration: Successfully integrated Label Knowledge Graphs with Multimodal Chain-of-Thought reasoning.
- Outcome: State-of-the-Art (SOTA) results on standard benchmarks (NUS-WIDE and Open Images).

Impact:

- The approach mimics human cognitive processes (Observe → Reason → Conclude).
- Explicit capturing of label co-occurrences improves consistency in multi-label predictions.
- Proven effectiveness for handling "unseen" concepts in dynamic real-world applications (e.g., image annotation).

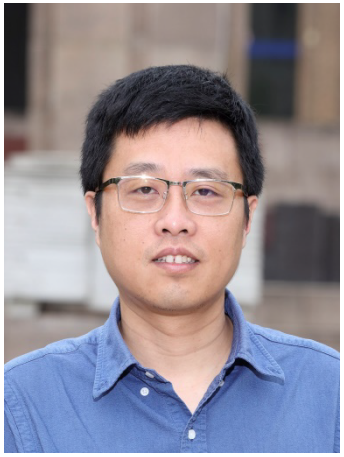


Xiang WEN received his MS degree from Beijing Jiaotong University. He is currently pursuing his PhD degree in Zhejiang University. His research interests include user profiling, deep learning, and artificial intelligence applications across online games.

Haobo WANG received his BS degree in computer science and technology from Zhejiang University in 2018. He is currently pursuing his PhD degree at the College of Computer Science and Technology, Zhejiang University. His research interests include machine learning and data mining, especially on weakly-supervised learning and multi-label learning.

Ke CHEN received her PhD degree in computer science from Zhejiang University in 2007. She is an associate professor with the College of Computer Science and Technology, Zhejiang University. Her research interests include database, large-scale data management technologies, and data privacy protection.

Tianlei HU received his PhD degree in computer science from Zhejiang University. He is an associate professor with the College of Computer Science and Technology, Zhejiang University. His research interests include large-scale data management and mining technologies supporting massive Internet users.



Gang CHEN received his PhD degree in computer science from Zhejiang University. He is a professor with the College of Computer Science and Technology, Zhejiang University. He is the Director of the Key Laboratory of Intelligent Computing Based Big Data of Zhejiang Province. He is a member of IEEE and ACM, and a standing member of the Database Professional Committee of the China Computer Federation. His research interests include database management technology, intelligent computing based big data, and massive Internet systems.