Exploiting Descriptive Completeness Prior for Cross Modal Hashing with Incomplete Labels

Haoyang Luo[1], Zheng Zhang[1], Yadan Luo[2]

[1]Harbin Institute of Technology, Shenzhen [2]University of Queensland

NeurIPS 2024 poster paper

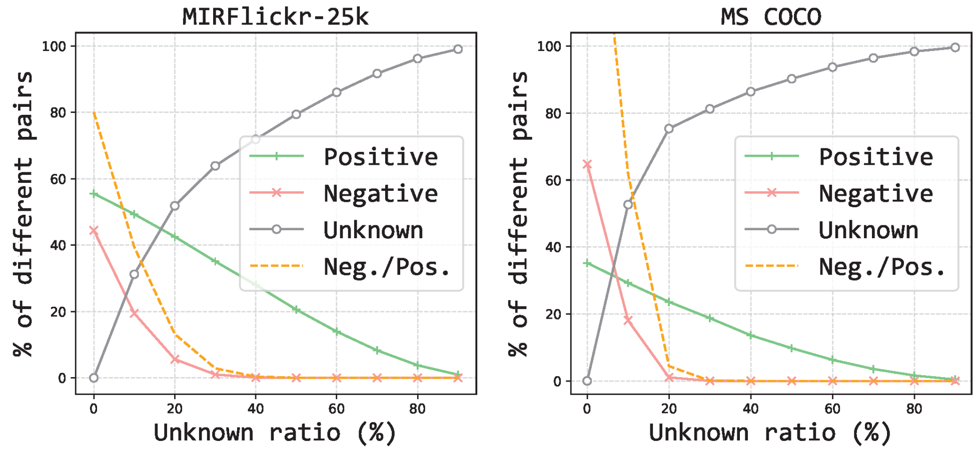
Motivation

Cross-modal hashing (CMH) addresses the highly demanding cross-modal similarity search in both web search systems and academic domains.

- However, due to limited labour resources, fully supervised annotation becomes impractical for large-scale datasets. **Partial annotation** with unknown labels is a feasible solution for Multi-label learning systems.

CMH with partial labels inevitably encounters disrupted similarity learning.

- Jointly missing labels across samples can produce:
  - ➢ Reduced positive pairs.
  - ➢ Unclear relationship between negative pairs.
- Existing CMHs require clear supervision from pairwise similarity



Motivation

Some feasible solution leveraging prior knowledge in **vision-language models** (e.g., CLIP) has been established for the partial multi-label recognition task.

However, CLIP label recovery for deep CMH remains under-explored because the original CLIP prompt yields an unsatisfactory 68% recovery precision.

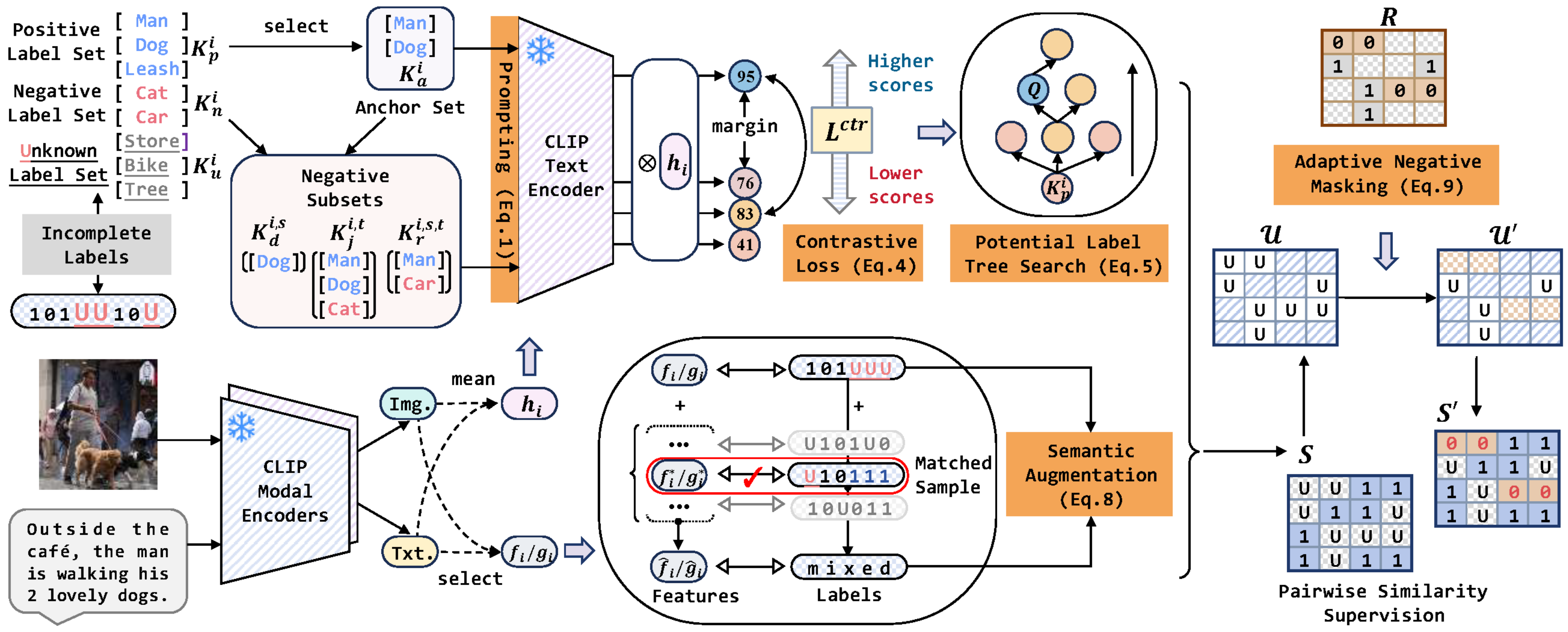
We seek to overcome the deficiencies of the original CLIP and consider the CLIP prior knowledge of **descriptive completeness**.

Contribution

- We propose a PCRIL framework, which jointly performs **semantic recovery** and **pairwise uncertainty elimination** for efficient cross-modal hashing with incomplete labels.
- A novel recovery architecture is proposed to recover the neglected semantic labels and pairwise similarities in the following figure.
- Extensive experiments verify that our PCRIL can consistently outperform state-of-the-art CMH methods across a range of incompleteness levels and different benchmarks.

Figure:

The General Framework.



Method – Prompt Contrastive Recovery

1. Contrastive Label Sets Construction.

For a sample i's positive labels

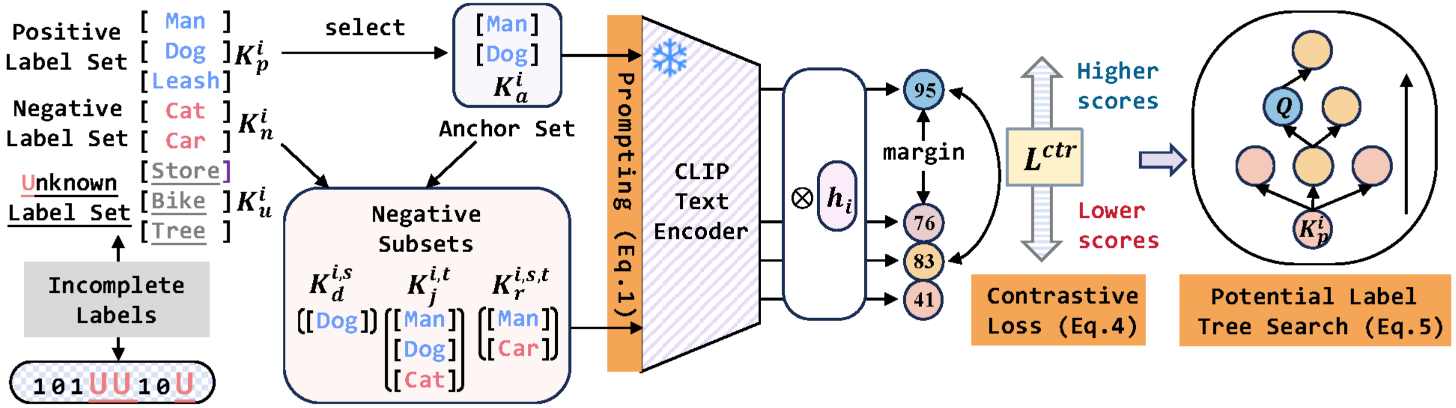
$$K_p^i = \{c \mid l_i^c = 1\},$$

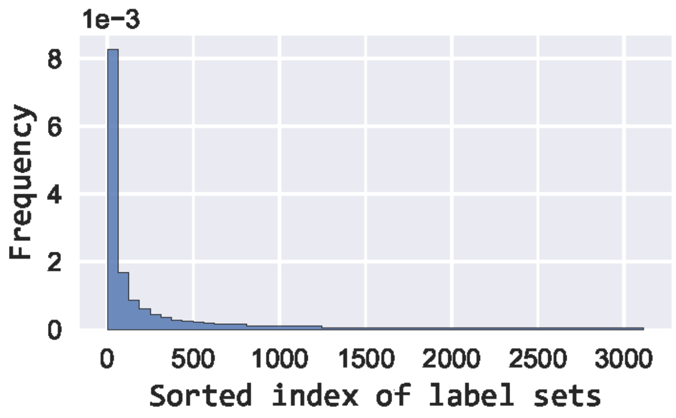
random subset $K_a^i \subset K_p^i$ is selected as the **anchor set**.

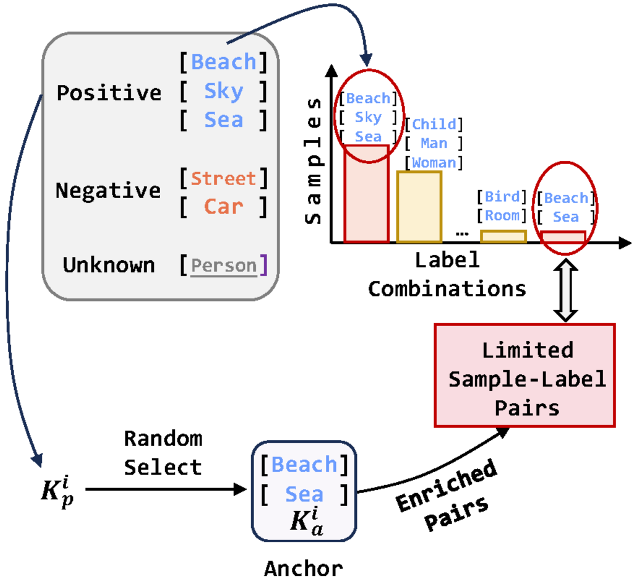
3 types of **negative variants** are constructed:

- deleting: $K_d^{i,s} = K_a^i - \{s\}$
- joining: $K_j^{i,t} = K_a^i \cup \{t\}$
- replacing: $K_r^{i,s,t} = K_a^i - \{s\} \cup \{t\}$

$$s \in K_a^i \quad t \in K_n^i$$







(a) The motivation of selecting positive anchor sets to enrich sample-label pairs.

Method – Prompt Contrastive Recovery

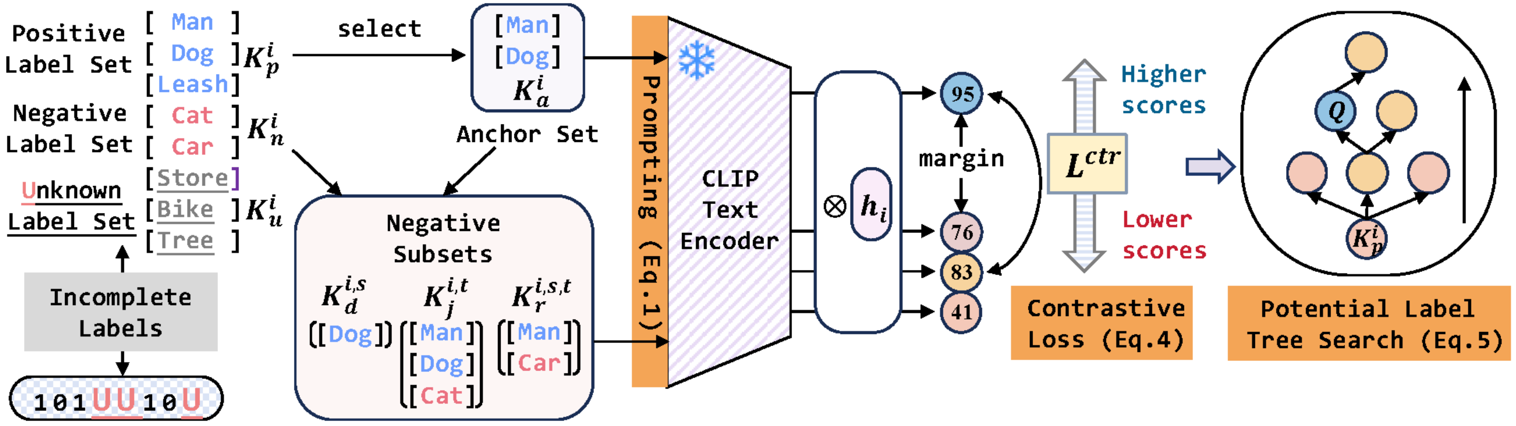
2. Prompt Contrastive Learning.

We define a **learnable prompt** template for multi-class sets as

$$\boldsymbol{P}(K_a^i) = (\boldsymbol{p}_{head}, \sigma(\{\boldsymbol{p}^c\}_{c\in K_a^i}), \boldsymbol{p}_{tail})$$

where $\boldsymbol{p}^c = (\boldsymbol{u}_1^c, \boldsymbol{u}_2^c, ..., \boldsymbol{u}_m^c, \mathrm{CLS}^c, \boldsymbol{v}_1^c, \boldsymbol{v}_2^c, ..., \boldsymbol{v}_n^c)$

contains learnable parameters.



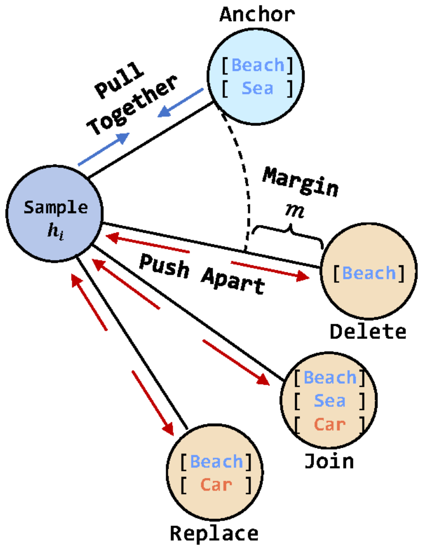
A CLIP matching score is defined as

$$\Phi^i(K) = E_t(P(K))^\top \boldsymbol{h}_i/\tau$$

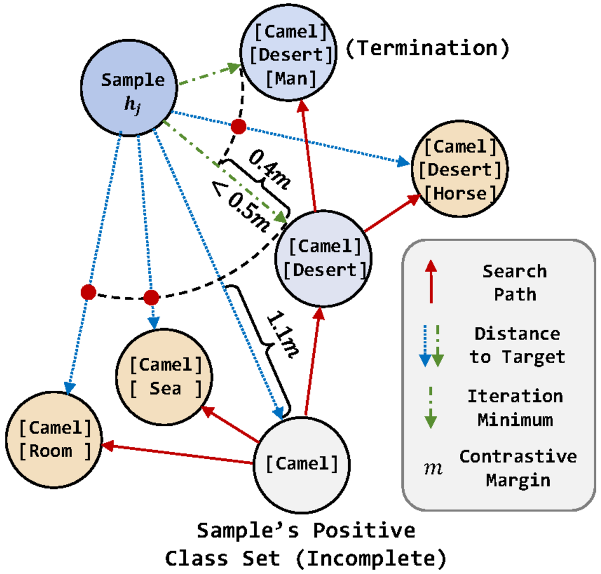
with the following **contrastive margin loss** between the anchor set and the 3 negative variants to learn completeness measurement:

$$\mathcal{L}^i(K_a, K_*) = max(\Phi^i(K_*) - \Phi^i(K_a) + m, 0)$$

$$\mathcal{L}^{ctr} = \sum_{i=1}^{N} \sum_{K_a^i \subset K_p^i} \Big(\sum_{s\in K_a^i} \mathcal{L}^i(K_a^i, K_d^{i,s}) + \sum_{t\in K_n^i} \mathcal{L}^i(K_a^i, K_j^{i,t}) + \sum_{s,t} \mathcal{L}^i(K_a^i, K_r^{i,s,t}) \Big)$$



(b) The relationship of the anchor and negative subsets.



(c) The potential label tree search with contrastively learned label embeddings.

Method – Prompt Contrastive Recovery

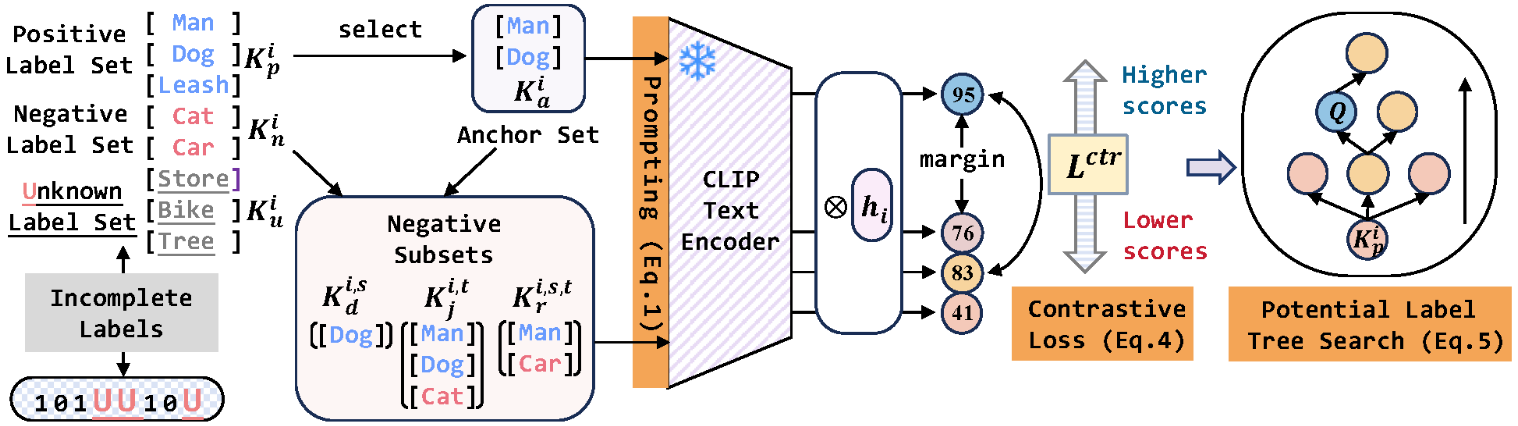
3. Potential Label Tree Search (PLTS).

After prompt contrastive learning, a tree-search process is defined as

$$c_u^* = \arg\max_{c_u \in K_u^i(\omega)} \Phi^i(K_p^i(\omega) \cup \{c_u\})$$

to search for a positive class in each step, with termination condition

$$\Phi^i(K_p^i(\omega^*) \cup \{c_u^*\}) < \Phi^i(K_p^i(\omega^*)) + \frac{m}{2}$$

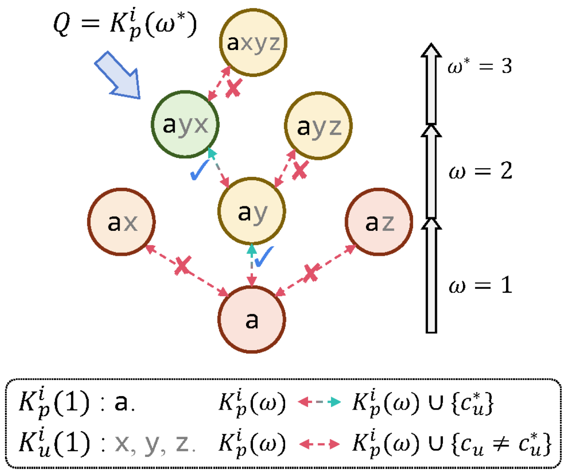


We can further recover the remaining sample labels by specifying pseudo-labels:

$$l_i^{c_u} = H(\Phi^i(Q \cup \{c_u\}) - \phi)$$

where

$$H(x) = \max(0, \min(1, \tfrac{1}{2} + \tfrac{x}{m}))$$

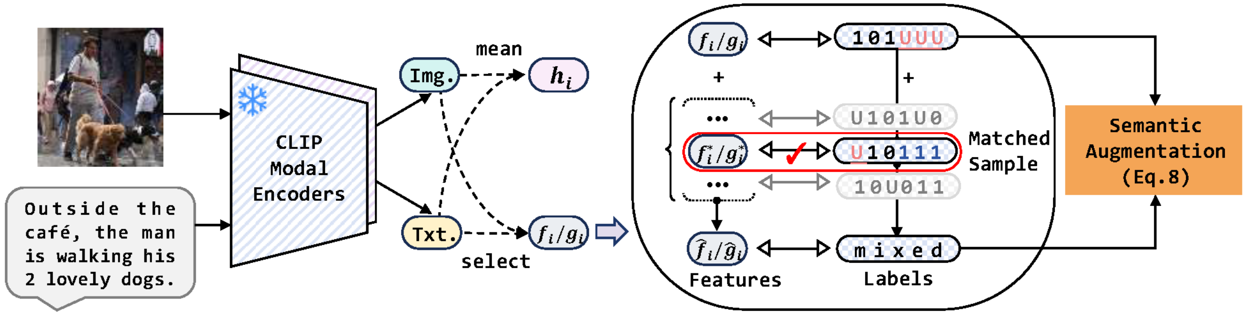


Method – Augmentation Strategies

1. Complementary Semantic Augmentation.

We mix up **complementary** samples to further eliminate uncertainty in labels.

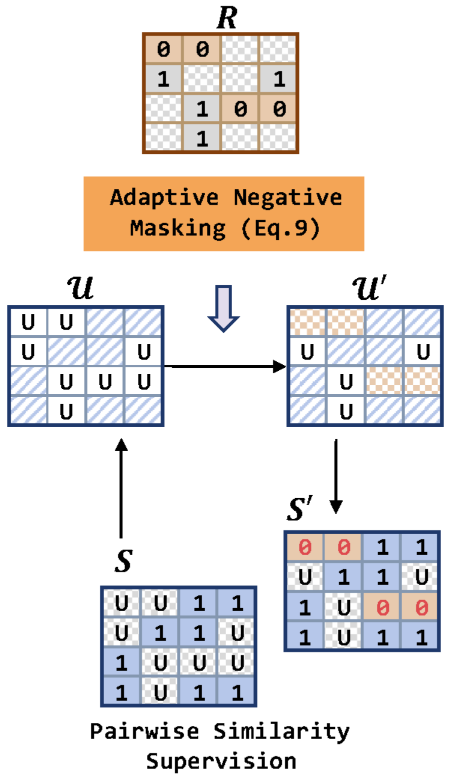
Samples carrying the same labels of respectively unknown and positive values are considered complementary.



2. Adaptive Negative Masking.

For pairwise supervision, we randomly flip a small proportion of unknown values (U) in similarity matrix as negative (0).

This prevents the false negative pairs from dominating the pairwise similarity learning.



Experimental Results

1. Cross-modal Hashing with Incomplete Labels.

Our method significantly remedies the current CMH methods for learning with incomplete labels.

Dataset	Method	30% known labels			50% known labels			70% known labels			Mean
		i→t	t→i	Mean	i→t	t→i	Mean	i→t	t→i	Mean	
Flickr	DCH [32]	69.8	65.9	67.8	75.7	70.2	72.9	77.5	72.1	74.8	71.9
	SDMCH [26]	64.3	67.2	65.8	66.0	73.9	70.0	69.5	76.0	72.8	69.5
	SCRATCH [3]	75.8	68.7	72.2	82.1	74.6	78.3	85.0	77.8	81.4	77.3
	WCHash [22]*	-	-	-	-	-	-	62.5	62.6	62.6	-
	DCMH [14]	63.0	65.2	64.1	67.4	70.2	68.8	71.3	74.5	72.9	68.6
	SSAH [17]	58.8	67.6	63.2	69.2	73.3	71.3	75.3	77.4	76.4	70.3
	AGAH [11]	59.8	63.4	61.6	78.4	76.6	77.5	84.1	79.2	81.6	73.6
	DCHMT [30]	64.1	64.0	64.0	78.3	75.6	76.9	81.0	80.0	80.5	73.8
	PCRIL (ours)	**78.5** (2.7)	**75.4** (6.7)	**77.0** (4.8)	**85.4** (3.3)	**79.4** (2.8)	**82.4** (4.1)	**87.5** (2.5)	**82.2** (2.2)	**84.9** (3.3)	**81.4** (4.1)
NUS	DCH [32]	65.1	66.1	65.6	65.2	66.9	66.0	67.1	68.2	67.6	66.4
	SDMCH [26]	55.7	59.9	57.8	58.9	61.2	60.0	59.3	62.2	60.7	59.5
	SCRATCH [3]	35.5	64.1	49.8	28.9	67.4	48.2	32.6	68.9	50.7	49.6
	DCMH [14]	29.5	31.3	30.4	32.4	33.4	32.9	36.3	35.5	35.9	33.1
	SSAH [17]	35.9	45.3	40.6	38.4	57.1	47.8	46.7	64.0	55.3	47.9
	AGAH [11]	46.7	49.7	48.2	58.8	49.9	54.4	66.7	67.2	66.9	56.5
	DCHMT [30]	35.7	35.0	35.4	57.6	55.9	56.7	67.3	67.4	67.4	53.1
	PCRIL (ours)	**67.2** (2.1)	**70.1** (4.0)	**68.7** (3.1)	**68.9** (3.7)	**70.4** (3.0)	**69.7** (3.7)	**70.4** (3.1)	**72.3** (3.4)	**71.4** (3.8)	**69.9** (3.5)
COCO	DCH [32]	60.9	61.1	61.0	63.0	63.4	63.2	64.2	64.9	64.5	62.9
	SDMCH [26]	53.7	55.5	54.6	57.3	56.9	57.1	58.5	58.7	58.6	56.8
	SCRATCH [3]	33.5	59.1	46.3	34.6	60.9	47.8	32.6	63.4	48.0	47.4
	DCMH [14]	49.2	47.0	48.1	52.3	53.1	52.7	52.9	53.1	53.0	51.3
	SSAH [17]	32.0	40.4	36.2	31.1	50.5	40.8	36.7	55.6	46.1	41.0
	AGAH [11]	54.2	56.1	55.1	58.5	58.8	58.6	61.2	62.4	61.8	58.5
	DCHMT [30]	44.8	44.3	44.5	52.1	49.5	50.8	62.0	61.5	61.8	52.4
	PCRIL (ours)	**62.8** (1.9)	**63.5** (2.4)	**63.2** (2.2)	**64.0** (1.0)	**64.7** (1.3)	**64.4** (1.2)	**67.8** (3.6)	**68.8** (3.9)	**68.3** (3.8)	**65.3** (2.4)

Experimental Results

2. Ablation Study.

Each contribution has stably improved the performance:

- ANM significantly outperforms traditional settings such as Assume Negative (AN), verifying our balanced similarity supervision.
- PCR and CSA modules perform reliable label recovery and improve performance through all datasets and settings.

Method	Flickr			NUS			COCO		
	30% known	50% known	70% known	30% known	50% known	70% known	30% known	50% known	70% known
B w/ IU [9]	57.5	73.4	82.8	62.4	63.3	67.5	49.6	50.4	45.9
B w/ AN [6]	68.9	76.6	81.5	51.1	53.8	66.2	45.8	54.3	59.8
B w/ ANM	75.0	78.1	83.8	60.6	60.7	68.1	59.9	61.4	65.1
B w/ ANM + PCR	76.3	82.1	84.4	68.0	69.4	70.9	62.4	63.7	67.2
B w/ ANM + PCR + CSA	**77.0**	**82.4**	**84.9**	**68.7**	**69.7**	**71.4**	**63.2**	**64.4**	**68.3**

Method	Flickr			COCO		
	30% known	50% known	70% known	30% known	50% known	70% known
B w/ AN	68.9	76.6	81.5	45.8	54.3	59.8
B w/ AN + CSP	68.8	76.2	82.3	46.4	54.1	59.7
B w/ AN + PCR	75.0	79.4	82.9	55.7	58.2	**65.8**
B w/ AN + PCR + CSA	**75.3**	**80.2**	**83.5**	**58.6**	**59.6**	65.2

Experimental Results

3. Prompt construction and recovery.

- Our prompt construction outperforms the conventional single-label prompts and the pure textual prompts.

- The tree-search label recovery sheme PLTS generally produces the best results compared to one-step labeling and single-modal conditioning strategies.

Table 4: Prompt construction variants compared on Flickr dataset. The MAP and precisions of recovered positive labels (PRECISION) are reported. Our PCRIL can successfully marry multi-label information with CLIP prior knowledge (compared to Conventional) and yield learned prompts for instance-label matching (compared to Phrasal).

Variant	Prompt Type		MAP				PRECISION			
	Learnable	Multi-label	30% known	50% known	70% known	Mean	30% known	50% known	70% known	Mean
Phrasal		✓	75.0	76.9	74.0	75.3	65.5	68.2	68.3	67.3
Conventional	✓		76.3	81.8	82.8	80.3	86.0	89.6	87.0	87.5
Ours	✓	✓	**77.0**	**82.4**	**84.9**	**81.4**	**87.4**	**89.6**	**92.0**	**89.7**

Table 5: Prompt search variants compared on Flickr and NUS datasets. Compared to single-modal recovery, our proposed PLTS can perform instance-level matching to produce more precise results. The one-step all variant validates the effectiveness of our recursive label recovery in PLTS.

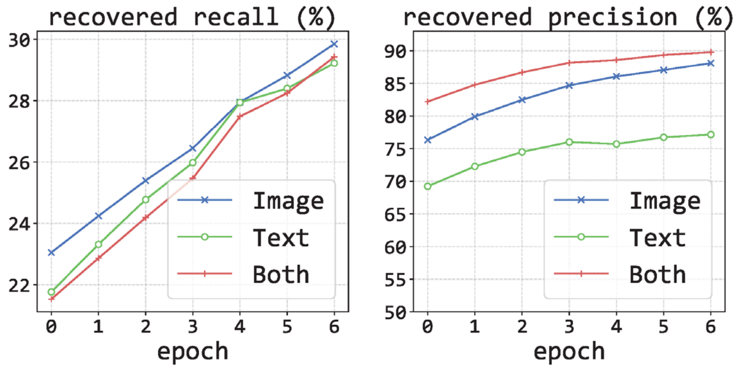
Dataset	Variant	MAP				PRECISION			
		30% known	50% known	70% known	Mean	30% known	50% known	70% known	Mean
Flickr	By image	**78.2**	79.2	**85.3**	80.9	86.2	86.6	88.1	87.0
	By text	74.3	76.6	84.4	78.4	70.1	80.2	77.2	75.8
	One-step all	64.6	77.4	82.9	75.0	21.0	38.4	54.6	38.0
	Ours	77.0	**82.4**	84.9	**81.4**	**87.4**	**89.6**	**92.0**	**89.7**
NUS	By image	51.3	65.4	68.5	61.7	78.4	76.0	74.9	76.4
	By text	50.8	65.1	69.3	61.7	69.4	69.1	68.9	69.1
	One-step all	48.4	64.9	67.5	60.2	12.1	23.7	27.1	21.0
	Ours	**68.7**	**69.7**	**71.4**	**69.9**	**79.5**	**78.1**	**80.3**	**79.3**

Experimental Results

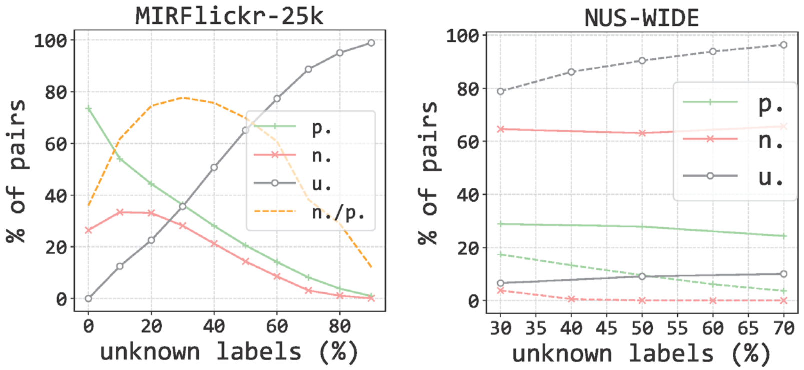
4. Other quantitative & qualitative results also verifies the effectiveness of our method from various aspects.

a) Label recovery through epoch. b) Pairwise supervision recovery. c) Feature space: baseline vs. recovered.

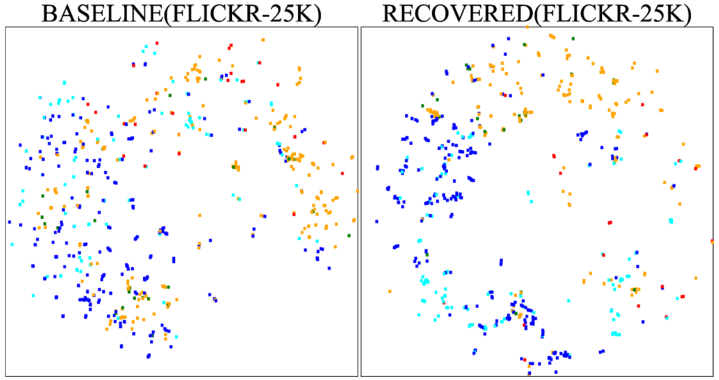
b) d) PLTS visualization. e) Heatmap for CLIP recovery model.



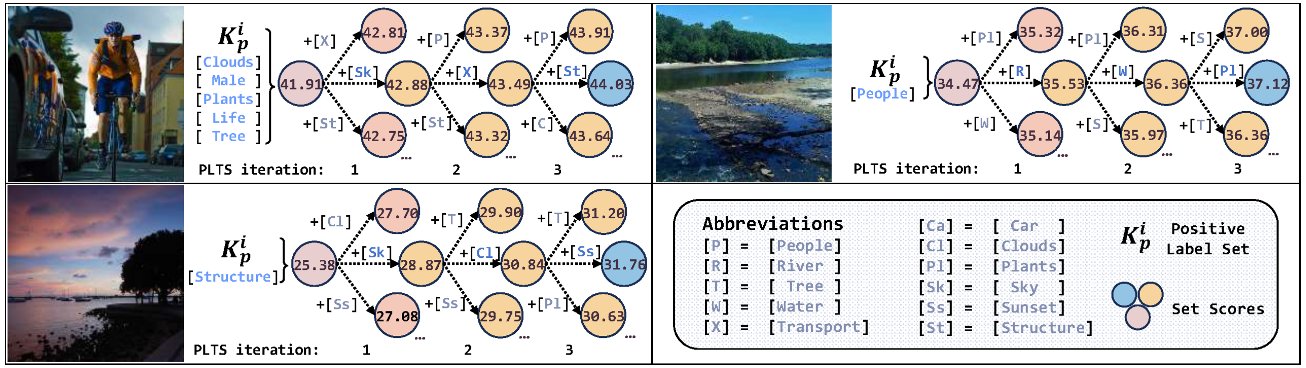
a)



b)



c)



d)

e)

Conclusion

- We propose a PCRIL framework, which jointly performs semantic recovery and pairwise uncertainty elimination for efficient cross-modal hashing with incomplete labels.
- To the best of our knowledge, this is the first CMH method to enable prompt learning with incomplete labels..
- Extensive experiments on widely used benchmarks validated that PCRIL can significantly outperform state-of-the-art CMH methods with different partial levels.