PCoRT: Progressive Chain-of-Thought Retrieval Tuning for Robust Semi-Supervised Vision

Anonymous authors
Paper under double-blind review

Abstract

Semi-supervised learning (SSL) typically exploits unlabeled data through pseudo-labeling, but predictions on external or heterogeneous data can be unreliable when their distributions differ from that of the labeled target data. Recent work has introduced Chain-of-Thought (CoT) reasoning to provide additional semantic information for unlabeled visual data. However, existing approaches generally treat a CoT as a single semantic unit, leaving the distinct information conveyed by its intermediate stages underused. We propose **Progressive Chain-of-Thought Retrieval Tuning (PCoRT)**, which preserves the stage-wise structure of CoT and progressively retrieves and fuses stage-specific semantic evidence into visual representations. PCoRT supports two settings with different uses of unlabeled data: in the in-distribution setting, CoT-enhanced predictions refine conventional pseudo-supervision; in the out-of-distribution setting, external unlabeled data provide auxiliary semantic evidence rather than direct pseudo-labels for the target task. We further distill the CoT-enhanced teacher into a visual-only student, which does not require the auxiliary CoT, retrieval, or fusion modules at deployment. Experiments on image classification, object detection, and semantic segmentation across four unlabeled-data settings show that PCoRT achieves competitive performance in all settings, with particularly strong gains in cross-dataset and heterogeneous multi-source settings.

1 Introduction

Semi-supervised learning (SSL) typically exploits unlabeled data by assigning pseudo-labels based on model predictions (Sohn et al., 2020; Zhang et al., 2021; Wang et al., 2022b; Chen et al., 2023a). When the unlabeled data are sufficiently aligned with the labeled target data, such pseudo-supervision can provide useful training signals (Oliver et al., 2018). In practice, however, unlabeled data may come from external or heterogeneous sources that differ from the target distribution. Predictions on such data may be unreliable, making direct pseudo-labeling less effective (Oliver et al., 2018; Saito et al., 2021; Huang et al., 2021). Distribution mismatch does not necessarily render external data useless, however: these data may still contain semantic information relevant to the target task. This raises an important question: how can SSL benefit from the semantic content of external unlabeled data when their task predictions cannot be reliably used as direct supervision?

Recent work has begun to incorporate Chain-of-Thought (CoT) reasoning into SSL, using language-based reasoning to provide additional semantic information for unlabeled images. For example, CERS (Chen et al., 2026) constructs a reasoning knowledge pool and retrieves relevant CoT information during training. This line of work offers a promising way to exploit semantic information, but existing methods typically encode and retrieve a complete CoT as a single semantic unit. As a result, they do not explicitly distinguish the information conveyed at different reasoning stages. A CoT may contain semantic cues at multiple levels, ranging from global context to object semantics, region relations, and task-relevant details (Divvala et al., 2009; Lu et al., 2016; Anderson et al., 2018). Collapsing these stages into a single representation may obscure their individual contributions and potential complementarity. We therefore investigate whether preserving the stage structure of CoT and progressively retrieving and integrating stage-specific cues can provide more effective semantic guidance for visual learning.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address these questions, we propose **Progressive Chain-of-Thought Retrieval Tuning (PCoRT)**, a framework for semi-supervised visual learning. PCoRT organizes CoT into a sequence of semantic stages and constructs stage-specific memories from unlabeled data. Given a visual representation, it progressively retrieves and fuses relevant evidence from these memories to produce a CoT-enhanced teacher. We consider two representative settings that differ in how the retrieved information is used. In the *in-distribution* setting, CoT-enhanced predictions refine conventional pseudo-supervision. In the *out-of-distribution* setting, evidence retrieved from external unlabeled data provides auxiliary semantic guidance, while predictions on those data are not directly used as pseudo-labels for the target task. Since the teacher requires additional CoT memory, retrieval, and multimodal fusion components during training, we further distill its task-specific knowledge into a visual-only student (Hinton et al., 2015). The resulting student can be deployed without the auxiliary CoT, retrieval, or fusion modules.

We evaluate PCoRT on image classification, object detection, and semantic segmentation across four unlabeled-data settings covering both the *in-distribution* and *out-of-distribution* scenarios. PCoRT achieves competitive performance in all settings across the evaluated tasks, with particularly notable gains in the more challenging cross-dataset and multi-source settings. On CIFAR-10 and CIFAR-100, PCoRT improves classification performance by up to 14.37% and 9.78%, respectively, while yielding gains of up to 4.8 mAP for object detection and 4.4 mIoU for semantic segmentation. Further experiments examine the effects of external data source, scale, and diversity. After distillation, the student retains a visual-only inference architecture and does not require the auxiliary CoT, retrieval, or fusion modules at deployment.

Our contributions are summarized as follows:

- We study the role of CoT reasoning structure in semi-supervised visual learning. Through comparisons with holistic CoT encoding, we evaluate the effect of preserving and progressively integrating stage-specific semantic cues.
- We propose **PCoRT**, which constructs stage-specific semantic memories and progressively retrieves and fuses reasoning evidence into visual representations. PCoRT supports both *in-distribution* and *out-of-distribution* settings by adapting how the retrieved evidence is used for supervision.
- We evaluate PCoRT on classification, detection, and segmentation across four unlabeled-data settings, demonstrating that it remains competitive across all settings. We further distill the CoT-enhanced teacher into visual-only students that can be deployed without the auxiliary reasoning modules.

2 Related Work

Semi-Supervised Learning under Distribution Shift. Semi-supervised learning (SSL) improves visual recognition by exploiting unlabeled data together with limited annotations. Mainstream methods, such as FixMatch (Sohn et al., 2020), FlexMatch (Zhang et al., 2021), FreeMatch (Wang et al., 2022b), and SoftMatch (Chen et al., 2023a), mainly combine pseudo-labeling with consistency regularization, and improve unlabeled-data utilization through confidence thresholds, sample weighting, or curriculum learning. Similar teacher–student and pseudo-labeling frameworks have also been widely studied for object detection and semantic segmentation (Liu et al., 2021; Zhang et al., 2023a; Wang et al., 2022c; Yang et al., 2023). However, these methods generally construct unlabeled supervision from task predictions generated by the model itself, which can become unreliable when labeled and unlabeled data follow different distributions. PCoRT explores a different direction by introducing language-based semantic reasoning to construct additional supervision for unlabeled data, providing a unified framework for classification, detection, and semantic segmentation.

LLM Reasoning for Vision. Chain-of-Thought (CoT) prompting enables large language models to solve complex tasks through intermediate reasoning steps (Wei et al., 2022). With the development of multimodal large language models, CoT has been further extended to visual reasoning, where intermediate textual rationales help connect visual observations with task predictions (Zhang et al., 2023b). Language-generated descriptions and rationales have also been used to provide auxiliary knowledge or supervision for visual models (Hsieh et al., 2023). Most existing studies, however, focus on using reasoning for answer generation, visual question answering, or knowledge distillation.

PCoRT instead focuses on semi-supervised visual learning: it organizes CoT into multiple semantic stages and uses the resulting reasoning to construct supervision for unlabeled images. The reasoning process is used only during training, while inference is performed by a standard visual model without additional language input.

3 METHODS

3.1 MOTIVATION

Unlabeled supervision across data regimes. Semi-supervised learning commonly uses predictions on unlabeled data as pseudo-supervision. When labeled and unlabeled data are sufficiently aligned, these predictions provide useful training signals. In practice, however, unlabeled data may come from external or heterogeneous sources. As the mismatch between these data and the target distribution increases, their task predictions become less reliable as direct supervision, as also observed in prior studies of semi-supervised learning under distribution mismatch (Oliver et al., 2018; Chen et al., 2020; Saito et al., 2021). Nevertheless, unreliable predictions do not make external data semantically irrelevant, as they may still contain useful contextual, object-level, relational, and task-related information. We therefore consider two regimes: *in-distribution*, where unlabeled predictions refine direct pseudo-supervision, and *out-of-distribution*, where external data provide auxiliary semantic evidence without their predictions being used as target-task pseudo-labels.

CoT as semantic guidance for semi-supervised learning. Recent work has introduced Chain-of-Thought (CoT) reasoning into semi-supervised learning. CERS (Chen et al., 2026) constructs a reasoning knowledge pool and employs CoT-aware retrieval to complement conventional visual supervision. This provides a promising mechanism for exploiting the semantic content of unlabeled data, particularly when their task predictions are not sufficiently reliable for direct pseudo-labeling. However, CoT is more than an additional textual description: it contains a sequence of intermediate semantic cues leading to a task-oriented conclusion. This raises a fundamental question of whether its internal stage structure should be preserved or collapsed into a single representation. CERS encodes the complete CoT as one embedding for retrieval and re-ranking, capturing its overall content without distinguishing the information conveyed at individual stages. Consequently, the contributions and potential complementarity of these stages remain underexplored.

From holistic reasoning to progressive integration. Preserving CoT structure raises the question of how to exploit information across reasoning stages. RAT (Wang et al., 2024b) iteratively retrieves evidence for intermediate thoughts conditioned on the current reasoning state and preceding context, illustrating the value of step-wise retrieval. Inspired by this principle, we investigate stage-wise retrieval and integration for visual CoT. Since visual recognition benefits from multi-level cues such as global context, object concepts, and relational information (Divvala et al., 2009; Anderson et al., 2018; Lu et al., 2016), we organize CoT into four stages: *global context*, *object semantics*, *region relations*, and *task relevant*. These stages provide potentially complementary evidence progressing from broad context to task-relevant details. Under the same CoT content, we compare holistic retrieval with structured progressive integration.

As shown in Fig. 1, we compare progressive structured CoT integration with holistic encoding (H) via linear probing on frozen visual features. Holistic encoding compresses all CoT stages into one representation, $f(c_1, \ldots, c_K)$, whereas structured CoT separately encodes and integrates each stage, $f(c_k)_{k=1}^{K}$, preserving stage-wise semantics. Progressive integration provides complementary information across visual recognition tasks and outperforms holistic encoding, showing that CoT's internal organization is important for effective semantic integration. This motivates Progres-

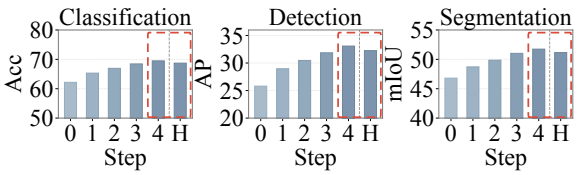

Figure 1: **Complementary information from structured CoT.** Step 0 is visual-only; Steps 1–4 progressively add *global-context, object-semantics, region-relations, and task-relevant*; H denotes holistic encoding of the complete CoT.

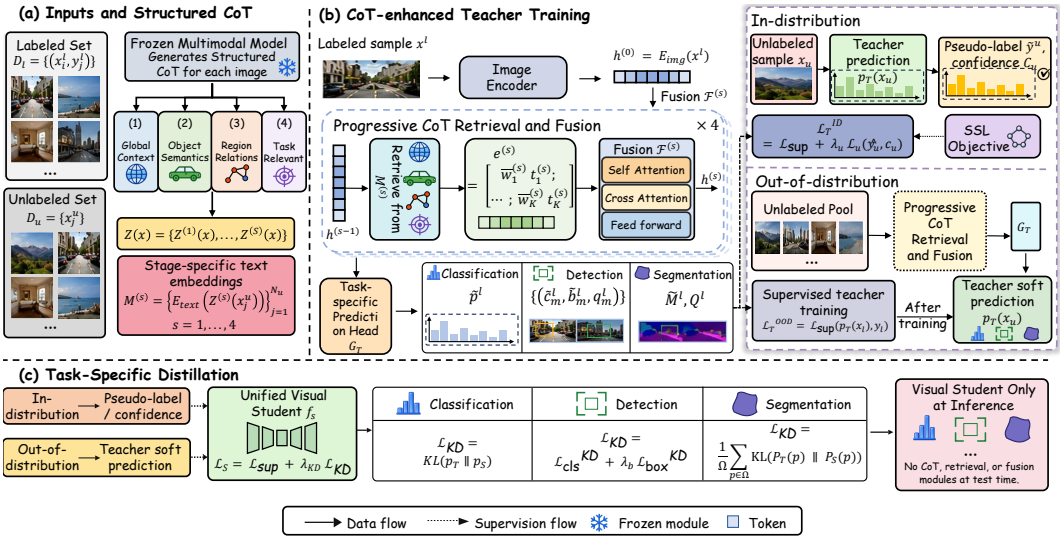

Figure 2: **Overview of the proposed PCoRT framework.** PCoRT performs structured CoT reasoning and retrieval-based supervision under two unlabeled-data regimes. For in-distribution, stage-wise CoT cues progressively refine pseudo-supervision targets for classification, detection, and semantic segmentation. For out-of-distribution unlabeled data, stage-specific CoT memories are retrieved and fused to enhance the teacher model. The resulting supervision is distilled into a unified visual student, while all CoT, retrieval, and fusion modules are removed at inference time.

sive Chain-of-Thought Retrieval Tuning (PCoRT), which builds stage-specific semantic memories and progressively retrieves and integrates the corresponding evidence.

Task-Specific Distillation. Progressive CoT Retrieval and Fusion enrich visual learning but introduce additional components that are unnecessary for deployment. Knowledge distillation provides an established way to transfer such training-time knowledge into a visual-only student without retaining the auxiliary modules at inference (Hinton et al., 2015). However, classification, detection, and segmentation differ substantially in their prediction structures, and prior work shows that effective distillation should be adapted accordingly (Wang et al., 2019; Liu et al., 2019). PCoRT therefore adopts task-specific distillation objectives to transfer the CoT-enhanced knowledge into the corresponding visual student, preserving the original visual-only architecture and introducing no additional inference-time cost.

3.2 Progressive Chain-of-Thought Retrieval Tuning

Based on the above motivation, we propose **Progressive Chain-of-Thought Retrieval Tuning (PCoRT)** (cf. Fig. 2), a general framework that uses structured CoT to improve semi-supervised learning. Instead of generating supervision only from the model's own predictions, PCoRT builds a semantic memory from structured CoT and retrieves relevant evidence to enhance the teacher model during training. The enhanced teacher is then combined with task-specific semi-supervised learning objectives and distilled into a visual-only student model. PCoRT supports both in-distribution and out-of-distribution semi-supervised settings by utilizing reasoning-enhanced supervision under different unlabeled data sources. The framework is applicable to classification, detection, and segmentation tasks, while introducing no additional cost during inference.

Problem Formulation and Data Regimes. We consider a general semi-supervised learning setting with a small labeled set and a larger unlabeled set. Let $\mathcal{D}_l = \{(x_i^l, y_i^l)\}_{i=1}^{N_l}$ denote the labeled dataset and $\mathcal{D}_u = \{x_j^u\}_{j=1}^{N_u}$ denote the unlabeled dataset, where $N_l \ll N_u$. The supervision target y varies according to the task. For classification, y represents the class label; for detection, y contains object categories and bounding boxes; for segmentation, y denotes the pixel-level semantic mask. PCoRT considers two types of unlabeled data regimes. In the **in-distribution** setting, labeled and unlabeled samples are drawn from the same data distribution, where the goal is to improve the quality of supervision generated from unlabeled examples. In the **out-of-distribution** setting, the unlabeled data may originate from sources different from those of the labeled data, leading to potential distribution gaps, requiring a more flexible way to transfer knowledge from unlabeled samples.

The following sections introduce how PCoRT constructs structured reasoning memory and utilizes it to enhance supervision under both settings.

Structured Multi-stage CoT. To construct structured semantic evidence from visual reasoning, PCoRT generates CoT from unlabeled images and represents each generated CoT as a structured multi-stage sequence rather than a single text description, denoted as $\mathbf{Z}(x) = \{Z^{(1)}(x), Z^{(2)}(x), Z^{(3)}(x), Z^{(4)}(x)\}$. Each element corresponds to one semantic stage, and we construct a separate memory for each stage to enable stage-conditioned retrieval. Specifically, the *global-context* stage captures the overall scene and high-level semantics. The *object-semantics* stage identifies the objects in the scene and captures their attributes and category-level meanings. The *region-relations* stage focuses on local structures and spatial interactions. The *task-relevant* stage provides information directly related to the prediction target, such as discriminative cues for classification, localization cues for detection, and boundary or region information for segmentation.

Each reasoning stage is encoded independently and stored in a corresponding semantic memory:

$$\mathcal{M}^{(s)} = \left\{ E_{\text{text}}\left(Z^{(s)}(x_j^u)\right) \right\}_{j=1}^{N_u}, \qquad s \in \{1,2,3,4\}. \tag{1}$$

where $Z^{(s)}(x_j^u)$ denotes the s-th reasoning stage generated for the unlabeled image x_j^u, and $E_{\text{text}}(\cdot)$ is the frozen text encoder. Maintaining stage-specific memories allows PCoRT to access different levels of semantic evidence progressively, rather than mixing all reasoning information into a single representation.

Progressive CoT Retrieval and Fusion. Given the stage-specific CoT memories, PCoRT progressively retrieves and fuses semantic evidence to enhance the visual representation. Specifically, the initial representation $h^{(0)} = E_{\text{img}}(x)$ is extracted from the input image and serves as the query for the first retrieval stage. At stage s, the current representation $h^{(s-1)}$ is used as the query to retrieve relevant CoT candidates from the corresponding memory:

$$\mathcal{C}^{(s)} = \text{TopK}\left(h^{(s-1)}, \mathcal{M}^{(s)}\right), \tag{2}$$

where $\mathcal{M}^{(s)}$ denotes the CoT memory of stage s. Each retrieved candidate is assigned a reliability score $w_k^{(s)} = \alpha_k^{(s)} \rho_k^{(s)}$, where $\alpha_k^{(s)}$ measures its relevance to the current visual representation and $\rho_k^{(s)}$ measures its visual consistency with the corresponding source image. The normalized reliability scores are then used to construct reliability-weighted semantic tokens:

$$\bar{w}_k^{(s)} = \frac{w_k^{(s)}}{\sum_j w_j^{(s)} + \epsilon}, \qquad e^{(s)} = \left[\bar{w}_1^{(s)} c_1^{(s)}; \ldots; \bar{w}_K^{(s)} c_K^{(s)}\right] \in \mathbb{R}^{K \times d}, \tag{3}$$

where $c_k^{(s)}$ denotes the embedding of the k-th retrieved CoT candidate. The resulting semantic evidence is integrated with the current visual representation through an ALBEF-style (Li et al., 2021) cross-modal attention block, yielding $h^{(s)} = \mathcal{F}(h^{(s-1)}, e^{(s)})$. The updated representation is subsequently used as the query for the next reasoning stage, forming a progressive process $h^{(0)} \rightarrow h^{(1)} \rightarrow \cdots \rightarrow h^{(S)}$, where $S = 4$ denotes the number of reasoning stages. Through this progressive retrieval and fusion process, PCoRT gradually enriches the visual representation with complementary semantic evidence from different reasoning stages.

CoT-enhanced Teacher Construction. After progressive CoT retrieval and fusion, the enhanced representation $h^{(S)}$ is fed into the task-specific prediction module to construct the CoT-enhanced teacher:

$$f_T^{\text{PCoRT}}(x, \mathcal{M}_u) = G_T(h^{(S)}), \tag{4}$$

where $G_T(\cdot)$ denotes the task-specific prediction module and $\mathcal{M}_u$ represents the unlabeled CoT memory. By incorporating retrieved semantic evidence into the visual representation, the PCoRT teacher produces stronger task predictions while preserving the original task architecture.

For the **in-distribution** regime, PCoRT follows the conventional pseudo-label-based semi-supervised learning paradigm. The CoT-enhanced teacher predicts unlabeled samples with retrieved

semantic evidence and generates pseudo-labels $\hat{y}_u$ and confidence scores c_u. Together with supervised learning on labeled samples, the teacher is optimized with the standard SSL objective:

$$\mathcal{L}_T^{ID} = \mathcal{L}_{sup} + \lambda_u \mathcal{L}_u(\hat{y}_u, c_u), \tag{5}$$

where $\mathcal{L}_{sup}$ denotes the supervised loss on labeled samples and $\mathcal{L}_u$ represents the task-specific unlabeled objective. Therefore, PCoRT does not modify the SSL optimization strategy, but improves pseudo-label reliability by enhancing the teacher representation.

For the **out-of-distribution** regime, the unlabeled samples are not directly used for pseudo-label optimization because distribution mismatch makes their target-task predictions unreliable. Instead, they provide semantic evidence through the unlabeled CoT memory $\mathcal{M}_u$. The CoT-enhanced teacher is trained on labeled images while leveraging semantic evidence retrieved from the unlabeled CoT memory $\mathcal{M}_u$. The teacher is then optimized with supervised learning:

$$\mathcal{L}_T^{OOD} = \mathcal{L}_{sup}\left(p_T(x_l), y_l\right). \tag{6}$$

After training, the teacher predicts unlabeled samples as $p_T(x_u) = f_T^{\mathrm{PCoRT}}(x_u, \mathcal{M}_u)$, incorporating knowledge enhanced by retrieved unlabeled semantic evidence for subsequent student optimization.

Task-Specific Knowledge Distillation. After obtaining the CoT-enhanced teacher under different data regimes, PCoRT transfers the knowledge captured by teacher predictions into a visual-only student model through task-specific knowledge distillation:

$$\mathcal{L}_S = \mathcal{L}_{sup} + \lambda_{KD} \mathcal{L}_{KD}, \tag{7}$$

where the task-specific distillation loss is defined as:

$$\mathcal{L}_{KD} = \begin{cases} \mathrm{KL}(p_T \| p_S), & \text{classification,} \\ \mathcal{L}_{cls}^{KD} + \lambda_b \mathcal{L}_{box}^{KD}, & \text{detection,} \\ \dfrac{1}{|\Omega|} \displaystyle\sum_{p \in \Omega} \mathrm{KL}\big(P_T(p) \| P_S(p)\big), & \text{segmentation.} \end{cases} \tag{8}$$

where λ_{KD} controls the contribution of the distillation loss. For classification, $\mathbf{p}_T$ and $\mathbf{p}_S$ denote the teacher and student prediction distributions, respectively. For detection, $\mathcal{L}_{cls}^{KD}$ and $\mathcal{L}_{box}^{KD}$ denote the distillation terms applied to classification predictions and bounding-box regression outputs, respectively, and λ_b balances the two terms. For segmentation, $P_T(p)$ and $P_S(p)$ denote the pixel-wise prediction distributions at pixel location p, and Ω represents the set of image pixels. $\mathrm{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence.

The student retains only the visual backbone and task-specific prediction module, $\hat{y} = f_S(x)$, discarding all CoT-related components. Thus, PCoRT introduces structured semantic reasoning only during training to enhance teacher supervision, while maintaining efficient visual-only inference without additional computational overhead.

4 Experiments

4.1 Experimental Setup

Benchmarks and metrics. We evaluate PCoRT on classification (CIFAR-10/100 (Krizhevsky et al., 2009), top-1 accuracy), detection (COCO-Partial 10% (Lin et al., 2014), mAP), and segmentation (Pascal VOC 1/16 (Everingham et al., 2010), mIoU).

Unlabeled-data settings. We consider four settings spanning two unlabeled-data regimes. **Settings-I and II** correspond to the *in-distribution* regime, where the unlabeled samples remain closely aligned with the target data: **Setting-I** follows standard in-domain SSL, while **Setting-II** adopts a transductive (Chapelle et al., 2006) protocol by additionally including target test images as unlabeled training data, with their labels strictly hidden. In contrast, **Settings-III and IV** correspond to the *out-of-distribution* regime, where the unlabeled pool is collected outside the target dataset: **Setting-III** uses a single external unlabeled dataset, whereas **Setting-IV** combines multiple heterogeneous external sources. Across all settings, the labeled target dataset and the complete official evaluation set remain unchanged; only the unlabeled training pool varies.

Table 1: Classification results (Accuracy, %) on CIFAR-10 and CIFAR-100 under four settings.

Method	CIFAR-10				CIFAR-100			
	Setting-I	Setting-II	Setting-III	Setting-IV	Setting-I	Setting-II	Setting-III	Setting-IV
Labeled Only	75.78				49.90			
MixMatch (Berthelot et al., 2019b)	63.81 ↓11.97	68.84 ↓6.94	60.31 ↓15.47	51.63 ↓24.15	32.41 ↓17.49	30.85 ↓19.05	22.36 ↓27.54	27.60 ↓22.30
ReMixMatch (Berthelot et al., 2019a)	90.12 ↑14.34	94.30 ↑18.52	58.63 ↓17.15	57.41 ↓18.37	57.25 ↑7.35	56.83 ↑6.93	41.77 ↓8.13	46.60 ↓3.30
FixMatch (Sohn et al., 2020)	92.53 ↑16.75	85.19 ↑9.41	75.29 ↓0.49	63.65 ↓12.13	53.88 ↑3.98	54.58 ↑4.68	51.75 ↑1.85	53.25 ↑3.35
FlexMatch (Zhang et al., 2021)	95.03 ↑19.25	91.44 ↑15.66	69.56 ↓6.22	65.72 ↓10.06	60.06 ↑10.16	59.90 ↑10.00	47.28 ↓2.62	50.70 ↑0.80
FreeMatch (Wang et al., 2022b)	95.10 ↑19.32	**95.95** ↑20.17	74.02 ↓1.76	77.33 ↑1.55	62.02 ↑12.12	61.14 ↑11.24	45.64 ↓4.26	50.66 ↑0.76
SoftMatch (Chen et al., 2023a)	94.90 ↑19.12	95.05 ↑19.27	70.90 ↓4.88	73.49 ↓2.29	62.13 ↑12.23	61.61 ↑11.71	46.91 ↓2.99	54.29 ↑4.39
PCoRT (Ours)	**95.35** ↑19.57	95.47 ↑19.69	**89.66** ↑13.88	**88.91** ↑13.13	**63.26** ↑13.36	**62.69** ↑12.79	**61.53** ↑11.63	**59.44** ↑9.54

Table 2: Results on object detection on COCO-Partial 10% and semantic segmentation on Pascal VOC 1/16 under four settings.

(a) Object Detection (mAP %)					(b) Semantic Segmentation (mIoU %)				
Method	Setting-I	Setting-II	Setting-III	Setting-IV	Method	Setting-I	Setting-II	Setting-III	Setting-IV
Labeled Only	28.4				Labeled Only	45.1			
DINO (Teacher) (Zhang et al., 2022)	35.0 ↑6.6	35.4 ↑7.0	24.5 ↓3.9	18.2 ↓10.2	U^{2}PL (Wang et al., 2022c)	68.0 ↑22.9	68.3 ↑23.2	50.7 ↑5.6	41.5 ↓3.6
Omni-DETR (Wang et al., 2022a)	41.3 ↑12.9	41.9 ↑13.5	30.2 ↑1.8	24.1 ↓4.3	DDFP (Wang et al., 2024a)	75.0 ↑29.9	75.4 ↑30.3	52.7 ↑7.6	43.6 ↓1.5
Semi-DETR (Zhang et al., 2023a)	43.5 ↑15.1	43.7 ↑15.3	31.5 ↑3.1	25.8 ↓2.6	UniMatch V1 (Yang et al., 2023)	75.2 ↑30.1	75.8 ↑30.7	53.4 ↑8.3	43.1 ↓2.0
Sparse Semi-DETR (Shehzadi et al., 2024)	44.3 ↑15.9	44.5 ↑16.1	32.1 ↑3.7	26.5 ↓1.9	CorrMatch (Sun et al., 2024)	76.4 ↑31.3	75.9 ↑30.8	56.1 ↑11.0	42.8 ↓2.3
MixPL (Chen et al., 2023b)	44.6 ↑16.2	44.9 ↑16.5	32.2 ↑3.8	27.1 ↓1.3	BeyondPixels (Howlader et al., 2024)	77.3 ↑32.2	78.1 ↑33.0	53.9 ↑8.8	46.1 ↑1.0
STEP-DETR (Shehzadi et al., 2025)	**45.4** ↑17.0	45.1 ↑16.7	32.8 ↑4.4	27.7 ↓0.7	UniMatch V2 (Yang et al., 2025)	79.0 ↑33.9	79.4 ↑34.3	56.1 ↑11.0	46.8 ↑1.7
PCoRT (Ours)	44.7 ↑16.3	**45.3** ↑16.9	**35.9** ↑7.5	**32.5** ↑4.1	**PCoRT (Ours)**	**79.6** ↑34.5	**79.7** ↑34.6	**59.8** ↑14.7	**51.2** ↑6.1

4.2 Compared with State-of-The-Art Methods

PCoRT Shows Growing Advantages under Distribution Shift. As shown in Tables 1 and 2, PCoRT performs competitively across all settings, with larger gains under stronger distribution shifts. Under Settings I and II, retrieved structured reasoning complements visual pseudo-supervision; its benefits become more pronounced in the cross-dataset and multi-source Settings III and IV, where visual pseudo-labels are less reliable.For classification, Settings III and IV use STL-10 (Coates et al., 2011) and a 100k mixture of Caltech-101 (Fei-Fei et al., 2004), ImageNet (Deng et al., 2009), iNaturalist (Van Horn et al., 2018), STL-10, and complementary CIFAR data, respectively; PCoRT improves the strongest baselines by up to 14.37% and 9.78% on CIFAR-10 and CIFAR-100. For detection, Settings III and IV use Open Images (Kuznetsova et al., 2020) and a mixture of Pascal VOC (Everingham et al., 2010), Objects365 (Shao et al., 2019), and Open Images, respectively, yielding gains of 3.1 and 4.8 mAP on COCO-Partial 10%. For segmentation, Settings III and IV use COCO-Stuff (Caesar et al., 2018) and a mixture of COCO-Stuff, ADE20K (Zhou et al., 2017), Cityscapes (Cordts et al., 2016), and Mapillary Vistas (Neuhold et al., 2017), respectively, yielding gains of 3.7 and 4.4 mIoU on Pascal VOC 1/16. These results show that structured reasoning becomes increasingly valuable as visual pseudo-labels become less reliable, while consistent gains across classification, detection, and segmentation demonstrate the generality of PCoRT.

PCoRT Remains Effective across Diverse Cross-Dataset Shifts. Figure 3 evaluates PCoRT under the *out-of-distribution* regime across DTD (Cimpoi et al., 2014), SUN397 (Xiao et al., 2010), RESISC45 (Cheng et al., 2017), Retinopathy (Zhai et al., 2019), CLEVR-C (Johnson et al., 2017), and KITTI (Geiger et al., 2012). PCoRT outperforms the strongest baseline for most labeled–unlabeled pairs in Setting III, with a mean gain of 0.84%, and improves all six targets by 0.69%–2.36% in the multi-source Setting IV, with the largest gains on RESISC45 (2.36%), KITTI (1.85%), and DTD (1.56%). These results demonstrate its robustness to diverse and heterogeneous unlabeled data.

4.3 Ablation Study

The Impact of Different Components. We systematically investigate PCoRT through ablations on (i) the progressive CoT stages, (ii) textual evidence injection, and (iii) multimodal fusion strategies, isolating the contribution of each key design choice as follows.

❶ **Progressive CoT Stages.** Table 3 shows that global semantics yields substantial gains over the visual-only baseline, especially for classification, from 70.90% to 87.83% in Setting-III and 73.49% to 86.57% in Setting-IV. Object semantics further improves all three tasks, while region relations and task-specific cues bring additional gains, particularly for detection and segmentation. These results suggest that global semantics provides a strong anchor, while progressively finer object, region, and task-level reasoning contributes complementary information for more structured prediction tasks.

❷ **Textual Evidence Injection Strategies.** Table 4 compares different forms of textual evidence across the three tasks. Image captions consistently improve performance, showing that textual semantics complement visual features. Unstructured CoT brings further gains, indicating that richer

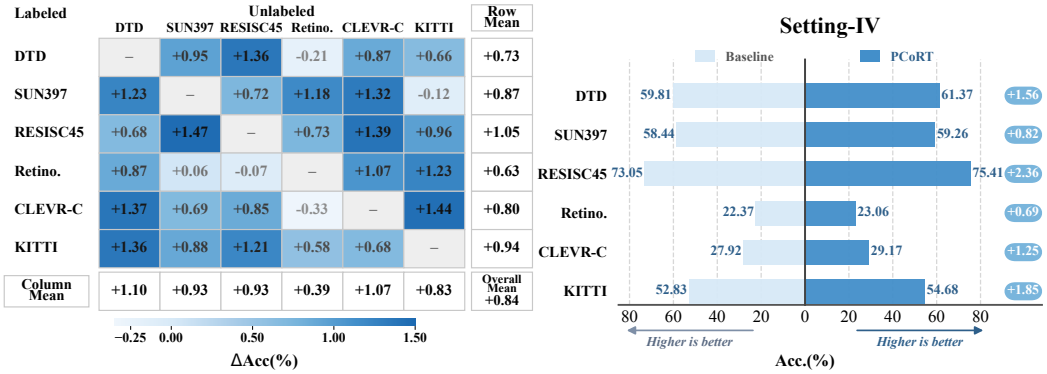

Figure 3: **PCoRT gains under distribution-shifted unlabeled data.** (a) Setting-III: pairwise gains between each labeled target and external unlabeled source. (b) Setting-IV: for each target, the other five datasets among DTD, SUN397, RESISC45, Retinopathy, CLEVR-C, and KITTI are merged as the unlabeled pool. Labeled and test data share the target dataset; $\Delta = \text{PCoRT} - \text{Best Baseline}$.

Table 3: **Ablation on progressive CoT stages.** Effects of progressively adding global semantics, object semantics, region relations, and task-specific cues under Settings III–IV.

Components / Ablations	Stage I	Stage II	Stage III	Stage IV	Classification Setting-III	Classification Setting-IV	Detection Setting-III	Detection Setting-IV	Segmentation Setting-III	Segmentation Setting-IV
exp-I (Visual only)	✗	✗	✗	✗	70.90	73.49	31.5	25.8	56.1	46.8
exp-II (+ Global Semantics)	✓	✗	✗	✗	87.83 (+16.93)	86.57 (+13.08)	33.1 (+1.6)	28.7 (+2.9)	56.9 (+0.8)	46.8 (+0.0)
exp-III (+ Object Semantics)	✓	✓	✗	✗	88.64 (+17.74)	87.46 (+13.97)	33.9 (+2.4)	30.0 (+4.2)	58.0 (+1.9)	48.2 (+1.4)
exp-IV (+ Region Relations)	✓	✓	✓	✗	89.18 (+18.28)	88.07 (+14.58)	35.0 (+3.5)	31.2 (+5.4)	58.9 (+2.8)	49.7 (+2.9)
exp-V (+ Task-specific Cues)	✓	✓	✓	✓	**89.66** (+18.76)	**88.91** (+15.42)	**35.9** (+4.4)	**32.5** (+6.7)	**59.8** (+3.7)	**51.2** (+4.4)

reasoning provides more informative semantic evidence. Structured CoT performs best across all tasks, showing that PCoRT benefits not only from textual information itself, but also from organizing it into structured reasoning stages.

❸ **Multimodal Fusion Strategies.** Table 5 compares strategies for integrating retrieved textual evidence. Residual and gated fusion outperform fixed-weight combination, while reliability-aware fusion brings further gains. Among non-progressive variants, cross-attention performs best across all three tasks, highlighting the importance of fine-grained cross-modal interaction. Progressive fusion achieves the strongest results by incorporating stage-specific evidence along the reasoning hierarchy rather than in a single step, with gains of 2.9% mAP and 3.5% mIoU over fixed-weight fusion for detection and segmentation under Setting-IV. These results show that effective CoT integration requires not only strong cross-modal fusion but also progressive semantic integration as visual representations evolve.

PCoRT Benefits from Increasing Unlabeled Scale and Diversity. Under a simulated out-of-distribution setting, we study the effects of unlabeled sample size (N_u) and category diversity (C_u), using **Caltech-101** and **CIFAR-100** as labeled targets and unlabeled data uniformly sampled from DTD, Flowers102 (Nilsback & Zisserman, 2008), Pets (Parkhi et al., 2012), SVHN (Netzer et al., 2011), and SUN397. PCoRT and the baseline use identical unlabeled pools, with the shaded regions in Fig. 4 indicating their performance gap.

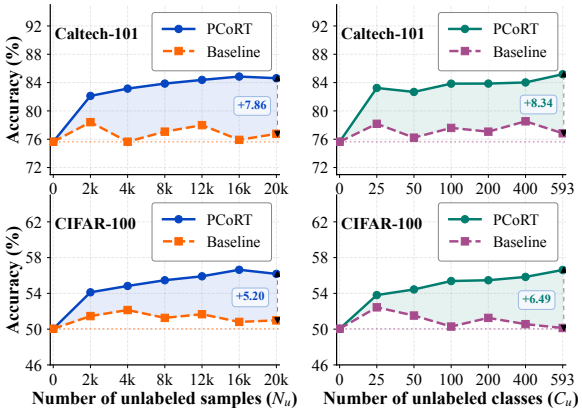

Figure 4: **Effect of unlabeled scale and diversity.** Performance of PCoRT and the baseline on Caltech-101 and CIFAR-100 as N_u (left) and C_u (right) vary.

❶ **Varying the Number of Unlabeled Samples.** To study unlabeled *scale*, we vary $N_u \in \{0, 2k, 4k, 8k, 12k, 16k, 20k\}$ with a fixed category space. As shown in the left column of Fig. 4, PCoRT benefits from increasing N_u, whereas the baseline fluctuates with limited improvement. At N_u=20k, the gap reaches 7.86%

Table 4: **Effect of textual semantics and CoT structure.** Comparison of no text, captions, unstructured CoT, and structured CoT under Settings III–IV across the three tasks.

Components / Ablations	Text	CoT	Structure	Classification		Detection		Segmentation	
				Setting-III	Setting-IV	Setting-III	Setting-IV	Setting-III	Setting-IV
exp-I (No Text)	✗	✗	✗	70.90	73.49	31.5	25.8	56.1	46.8
exp-II (Caption)	✓	✗	✗	87.45 (+16.55)	85.96 (+12.47)	34.6 (+3.1)	31.2 (+5.4)	57.8 (+1.7)	49.8 (+3.0)
exp-III (Unstructured CoT)	✓	✓	✗	88.32 (+17.42)	87.14 (+13.65)	35.1 (+3.6)	31.8 (+6.0)	58.6 (+2.5)	50.4 (+3.6)
exp-IV (Structured CoT)	✓	✓	✓	**89.66** (+18.76)	**88.91** (+15.42)	**35.9** (+4.4)	**32.5** (+6.7)	**59.8** (+3.7)	**51.2** (+4.4)

Table 5: **Effect of different fusion strategies.** We compare different fusion mechanisms under Setting-III and Setting-IV across classification, detection, and segmentation.

Fusion Strategy	Classification		Detection		Segmentation	
	Setting-III	Setting-IV	Setting-III	Setting-IV	Setting-III	Setting-IV
Fixed-weight	87.86	86.64	33.8	29.6	57.1	47.7
Residual	88.31 (+0.45)	87.14 (+0.50)	34.4 (+0.6)	30.3 (+0.7)	57.9 (+0.8)	48.7 (+1.0)
Gated	88.67 (+0.81)	87.69 (+1.05)	34.8 (+1.0)	30.8 (+1.2)	58.5 (+1.4)	49.3 (+1.6)
Reliability-aware	88.94 (+1.08)	88.03 (+1.39)	35.1 (+1.3)	31.4 (+1.8)	58.8 (+1.7)	49.9 (+2.2)
Cross-attention	89.21 (+1.35)	88.36 (+1.72)	35.4 (+1.6)	31.9 (+2.3)	59.2 (+2.1)	50.4 (+2.7)
Progressive Fusion	**89.66** (+1.80)	**88.91** (+2.27)	**35.9** (+2.1)	**32.5** (+2.9)	**59.8** (+2.7)	**51.2** (+3.5)

on Caltech-101 and 5.20% on CIFAR-100, but gradually saturates at larger budgets, suggesting that gains stem from richer semantic evidence rather than simply more unlabeled samples. A larger pool provides more candidates for stage-wise retrieval, increasing the chance of finding relevant and complementary reasoning cues, while reliability-aware selection suppresses mismatched evidence.

❷ **Varying the Number of Unlabeled Categories.** As shown in the right column of Fig. 4, we study unlabeled *diversity* by varying $C_u \in \{25, 50, 100, 200, 400, 593\}$ under a fixed unlabeled budget. The baseline saturates and fluctuates as C_u increases, whereas PCoRT shows an overall upward trend. With all 593 categories, PCoRT reaches 85.18% on Caltech-101 and 56.62% on CIFAR-100, exceeding the baseline by 8.34% and 6.49%. This shows that PCoRT benefits from semantic diversity even when many unlabeled categories do not match the target classes. By retrieving stage-specific global, object, region, and task-relevant semantics rather than pseudo-labeling such samples, PCoRT enriches its semantic memory while reducing sensitivity to heterogeneous label-space noise.

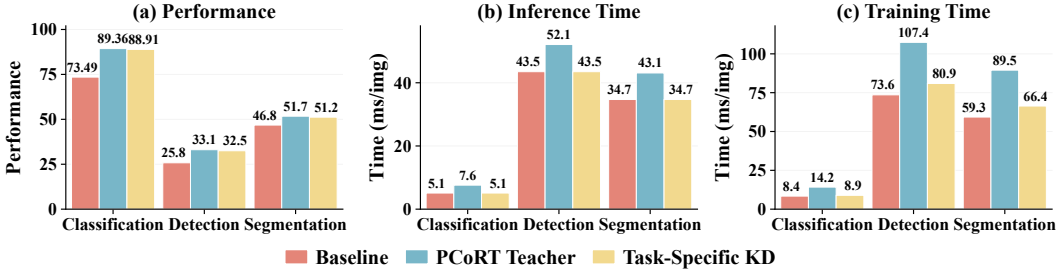

Figure 5: **Effect of task-specific distillation.** Comparison of the visual baseline, PCoRT teacher, and distilled student in performance, inference time, and training time across the three tasks.

Balancing Performance and Efficiency with Task-Specific Distillation. As shown in Fig. 5, the PCoRT teacher improves all three tasks but incurs extra overhead from reasoning, retrieval, and fusion. Task-specific distillation transfers these gains to a visual-only student, which remains within 0.45, 0.58, and 0.54 points of the teacher while restoring inference time to the baseline level with only minor additional training cost. Thus, distillation preserves nearly all PCoRT gains while enabling efficient visual-only inference.

5 Conclusion

We present **PCoRT**, a unified SSL framework for both *in-distribution* and *out-of-distribution* unlabeled data. By progressively retrieving and integrating structured CoT evidence, PCoRT improves unlabeled-data utilization under different distribution conditions. Extensive experiments across classification, detection, and segmentation demonstrate consistent improvements over SSL baselines.

AI USE STATEMENT

Generative AI tools were used only for language editing and improving the readability of the manuscript. They were not used to develop the methodology, formulate theoretical claims, design experiments, implement methods, or interpret results. All AI-assisted text was reviewed by the authors, and all technical content, claims, and reported results originate from our own work. The authors take full responsibility for the final content of the manuscript.

ETHICS STATEMENT

This work focuses on semi-supervised visual learning and is evaluated entirely on publicly available research datasets. We do not collect new personal data or conduct experiments involving human participants. As with other data-driven vision methods, the proposed framework may inherit biases or annotation artifacts from the underlying datasets and pretrained models. We therefore recommend appropriate validation before deployment in safety-critical or socially sensitive applications.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the experimental settings, datasets, evaluation metrics, structured CoT construction, stage-wise retrieval and fusion, and task-specific distillation objectives. The unlabeled-data protocols for all four settings are explicitly defined, and task-specific CoT prompts together with representative generations are provided in the appendix. We will release the implementation, prompts, and training configurations to facilitate reproduction of the reported results.

REFERENCES

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 6077–6086. IEEE, 2018.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019a.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019b.

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1209–1218. IEEE, 2018.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. A discussion of semi-supervised learning and transduction. In *Semi-supervised learning*, pp. 473–478. MIT Press, 2006.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. *arXiv preprint arXiv:2301.10921*, 2023a.

Yanbei Chen, Xiatian Zhu, Wei Li, and Shaogang Gong. Semi-supervised learning under class distribution mismatch. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3569–3576, 2020.

Yuming Chen, Yuxin Xie, Tao Zhou, and Yi Zhou. Beyond visual cues: Cot-enhanced reasoning for semi-supervised medical image segmentation. *arXiv preprint arXiv:2606.17958*, 2026.

Zeming Chen, Wenwei Zhang, Xinjiang Wang, Kai Chen, and Zhi Wang. Mixed pseudo labels for semi-supervised object detection. *arXiv preprint arXiv:2312.07006*, 2023b.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *2009 IEEE Conference on computer vision and Pattern Recognition*, pp. 1271–1278. IEEE, 2009.

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Prantik Howlader, Srijan Das, Hieu Le, and Dimitris Samaras. Beyond pixels: Semi-supervised semantic segmentation with a multi-scale patch-based multi-label classifier. In *European conference on computer vision*, pp. 342–360. Springer, 2024.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, 2023.

Junkai Huang, Chaowei Fang, Weikai Chen, Zhenhua Chai, Xiaolin Wei, Pengxu Wei, Liang Lin, and Guanbin Li. Trash to treasure: Harvesting ood data with cross-modal matching for open-set semi-supervised learning. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8290–8299. IEEE, 2021.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yen-Cheng Liu, Chih-Yao Ma, Zijian He, Chia-Wen Kuo, Kan Chen, Peizhao Zhang, Bichen Wu, Zsolt Kira, and Peter Vajda. Unbiased teacher for semi-supervised object detection. In *International Conference on Learning Representations*, 2021.

Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2604–2613, 2019.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pp. 852–869. Springer, 2016.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, pp. 7. Granada, 2011.

Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pp. 4990–4999, 2017.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Avital Oliver, Augustus Odena, Colin Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Kuniaki Saito, Donghyun Kim, and Kate Saenko. Openmatch: Open-set consistency regularization for semi-supervised learning with outliers. *arXiv preprint arXiv:2105.14148*, 2021.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8429–8438. IEEE, 2019.

Tahira Shehzadi, Khurram Azeem Hashmi, Didier Stricker, and Muhammad Zeshan Afzal. Sparse semi-detr: Sparse learnable queries for semi-supervised object detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5840–5850. IEEE, 2024.

Tahira Shehzadi, Khurram Azeem Hashmi, Shalini Sarode, Didier Stricker, and Muhammad Zeshan Afzal. Step-detr: Advancing detr-based semi-supervised object detection with super teacher and pseudo-label guided text queries. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3069–3079. IEEE, 2025.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3097–3107. IEEE, 2024.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, and Stefano Soatto. Omni-detr: Omni-supervised object detection with transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9357–9366. IEEE, 2022a.

Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4928–4937. IEEE, 2019.

Xiaoyang Wang, Huihui Bai, Limin Yu, Yao Zhao, and Jimin Xiao. Towards the uncharted: Density-descending feature perturbation for semi-supervised semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3303–3312. IEEE, 2024a.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022b.

Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4238–4247. IEEE, 2022c.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*, 2024b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7236–7246. IEEE, 2023.

Lihe Yang, Zhen Zhao, and Hengshuang Zhao. Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4): 3031–3048, 2025.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in neural information processing systems*, 34:18408–18419, 2021.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Jiacheng Zhang, Xiangru Lin, Wei Zhang, Kuo Wang, Xiao Tan, Junyu Han, Errui Ding, Jingdong Wang, and Guanbin Li. Semi-detr: Semi-supervised object detection with detection transformers. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23809–23818. IEEE, 2023a.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023b.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5122–5130. IEEE, 2017.

A Implementation Details

A.1 CoT Generation

We use Qwen2.5-VL-7B-Instruct to generate a four-stage CoT for each unlabeled image in a single inference call. The generator receives the image and task type, but is not provided with the dataset name, candidate classes, or ground-truth annotations. All generated CoTs are cached before training and remain fixed throughout the experiments. As illustrated in Fig. 6, we adopt a unified four-stage reasoning structure across classification, detection, and segmentation. Prompts and generated CoT examples are shown for the three tasks.

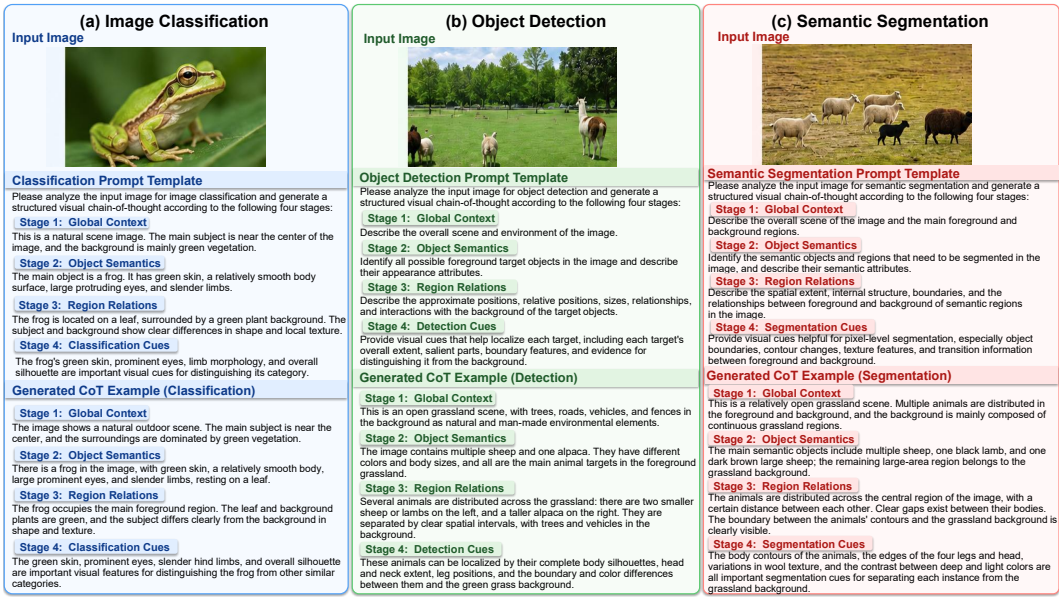

Figure 6: **Task-specific CoT prompts and representative generations.** We use a unified four-stage reasoning structure for classification, detection, and semantic segmentation.