DUALDISTILL: TRANSFERRING GENERATIVE RECOGNITION INTO EFFICIENT DISCRIMINATIVE PREDICTION FOR ANNOTATION-FREE MULTI-LABEL VISUAL RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

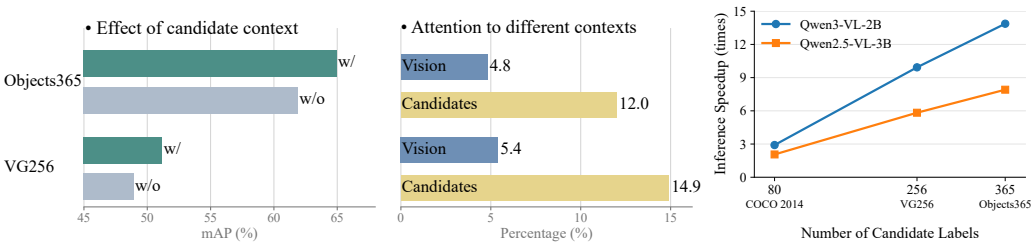
Traditional multi-label visual recognition relies heavily on human-annotated data, making it costly to scale to large datasets and label vocabularies. Multimodal Large Language Models (MLLMs) provide a promising alternative by performing annotation-free recognition through autoregressive label generation. However, this generative paradigm incurs substantial inference overhead due to task-prompt encoding and token-by-token decoding. In this paper, we investigate how to transfer the generative recognition capability of MLLMs into flexible and efficient discriminative prediction. We identify two key challenges in this transition: removing the task-aware prompt discards valuable candidate-label semantics, while directly optimizing MLLMs with a discriminative classification objective introduces a mismatch with their original generative training paradigm. To address these challenges, we propose DualDistill, a progressive generation-to-discrimination framework consisting of Cross-Paradigm Distillation (CPD) and Cross-Token Distillation (CTD). CPD first converts generative recognition into prompt-aware discriminative prediction while preserving the generative learning signal. CTD then transfers the resulting prompt-aware capability to an image-only pathway for efficient inference. Extensive experiments on multiple benchmarks demonstrate that DualDistill substantially improves both recognition performance and inference efficiency over strong MLLM fine-tuning and annotation-free baselines, achieving up to a 14 times inference speedup while maintaining strong classification performance without requiring any human annotations.

1 INTRODUCTION

Multi-label visual recognition aims to identify multiple semantic concepts present in a given image. It plays an important role in a wide range of real-world applications, including medical image analysis (Zhang et al., 2023), product tagging (Ferreira et al., 2021; Inoue et al., 2017), and remote sensing (Du et al., 2025). Despite substantial progress in classification performance, existing methods predominantly follow the conventional supervised learning paradigm, relying on human-annotated multi-label data for model training. Constructing such annotations requires considerable human effort and becomes increasingly costly when scaling to large datasets, expanding label vocabularies, or addressing domains that require specialized expertise. This heavy reliance on manual annotation therefore remains a major bottleneck to extending multi-label visual recognition to larger-scale datasets and broader real-world applications.

Foundation models provide a promising way to alleviate this annotation bottleneck by exploiting knowledge acquired from large-scale image-text data (Tan et al., 2024; Lin et al., 2025). Existing annotation-free approaches have primarily built on Vision-Language Models (VLMs), such as Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021; Lin et al., 2024), whose open-vocabulary recognition capability enables multi-label prediction without human annotations. More recently, Multimodal Large Language Models (MLLMs) have emerged as stronger visual recognizers that can directly identify multiple concepts through natural-language generation (Wang et al., 2024; Xie et al., 2026b). This makes MLLMs particularly attractive for annotation-free recognition,

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.



(a) Candidates context (b) Inference efficiency

Figure 1: Candidate-label context and inference efficiency. (a) Candidate-label context improves recognition and receives substantial attention from the prompt-side representation. (b) The image-side discriminative pathway achieves increasing inference speedups over autoregressive generation on datasets with larger label vocabularies, consistently across two MLLM backbones.

but their generative inference paradigm incurs substantial computational overhead: the model must process both the image and a task-aware prompt and then autoregressively generate labels token by token. Consequently, MLLMs remain considerably less computationally efficient than discriminative classifiers for large-scale or latency-sensitive visual recognition.

This raises an important question: can the generative recognition capability of an MLLM be progressively transferred into efficient discriminative prediction without losing the task-aware semantic context provided by the prompt? MLLM-generated annotations provide a natural source of supervision for such a transfer, enabling discriminative adaptation without human-annotated multi-label data. However, directly converting generative recognition into image-only discrimination is suboptimal due to two key gaps between the two paradigms. First, candidate-label context provides valuable task-specific semantics. As shown in Fig. 1(a), removing the candidate vocabulary consistently degrades recognition performance, while the prompt-side representation exhibits substantial attention to candidate-label tokens. However, under causal attention, the image-side classification token cannot access these prompt tokens, resulting in the loss of such semantic context. Second, replacing next-token prediction with a discriminative classification objective introduces an optimization mismatch with the original generative training paradigm of MLLMs, making classification-only adaptation less effective at exploiting their pretrained generative capability. These observations motivate a progressive transition from prompt-conditioned generation to efficient image-only discrimination.

To address these challenges, we propose DualDistill, a progressive generation-to-discrimination framework for efficient and annotation-free multi-label recognition. DualDistill establishes a prompt-aware discriminative pathway by introducing a classification token after the task-aware prompt. This token retains access to the candidate-label semantics used in generative recognition, while Generative-to-Discriminative (G2D) distillation converts teacher-generated responses into explicit discriminative supervision. To further alleviate the optimization mismatch between generation and discrimination, we preserve the teacher response in its original sequential form through a Generative-to-Generative (G2G) objective, retaining the next-token prediction signal during adaptation. Meanwhile, an image-side classification token provides a prompt-free inference pathway, and Cross-Token Distillation (CTD) transfers the richer task-aware knowledge from the prompt-side token to the image-side token. The resulting image-side classifier requires neither task-prompt encoding nor autoregressive decoding at inference time, enabling efficient single-forward prediction. As illustrated in Fig.1(b), this image-side pathway achieves increasingly pronounced inference speedups over autoregressive generation on datasets with larger label vocabularies, consistently across two MLLM backbones. Extensive experiments show that DualDistill substantially improves both recognition performance and inference efficiency over strong MLLM fine-tuning and annotation-free baselines, achieving up to a 14× inference speedup while maintaining strong classification performance.

2 PRELIMINARIES

We assume access to an unlabeled training set $\mathcal{D}_{\text{tr}} = \{\boldsymbol{x}_i\}_{i=1}^{n}$ and a predefined category vocabulary $\mathcal{C} = \{C_k\}_{k=1}^{q}$, while no image-level human annotations are available during training. Each image

$\boldsymbol{x}$ is associated with an unknown ground-truth multi-hot vector $\boldsymbol{y} \in \mathcal{Y}$, where $\mathcal{Y} = \{0,1\}^q$ is the label space. Here, $y_k = 1$ indicates that the k-th label with the name C_k is relevant while $y_k = 0$ indicates that it is not. Throughout this paper, $[\cdot;\cdot]$ denotes concatenation, while $\langle\cdot\rangle$ denotes a special token. Our goal is to fine-tune MLLMs on the training set $\mathcal{D}_{tr}$ without using any human image-level annotations, while improving both classification performance and inference efficiency on test images. We denote the frozen teacher MLLM and the trainable student MLLM as $\mathcal{M}^T$ and $\mathcal{M}^S$, respectively.

Generative MLLM for Multi-label Recognition We adopt the Qwen-VL family as the base MLLM, which consists of a vision encoder $\mathrm{Enc_v}(\cdot)$, a vision-language merger $\mathrm{Merger}(\cdot)$, and a language model composed of an LLM backbone $\mathrm{LLMBackbone}(\cdot)$ and a language modeling head $\mathrm{LMHead}(\cdot)$. Given an image $\boldsymbol{x}$ and a task-aware prompt $\mathcal{P}$ that specifies the recognition instruction, candidate label vocabulary, and output format, we obtain the visual and textual embeddings as $\mathbf{E}_\mathrm{v} = \mathrm{Merger}(\mathrm{Enc_v}(\boldsymbol{x}))$ and $\mathbf{E}_\mathrm{t} = \mathrm{Embed}(\mathcal{P})$, respectively, where $\mathrm{Embed}(\cdot)$ denotes the token embedding layer. For example, $\mathcal{P}$ can be: `What objects are in this image? Candidates: person, bicycle, ..., toothbrush. Please list only the objects, separated by commas.` These embeddings are jointly processed by the LLM for autoregressive generation. In conventional supervised fine-tuning, each image-prompt pair is associated with a target response containing the ground-truth labels, e.g., "person, dog". In our annotation-free setting, human-provided responses are unavailable; instead, we use responses generated by the frozen teacher $\mathcal{M}^T$ as supervision. Let $\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_T\}$ denote the tokenized target response. At decoding step j, the model predicts

$$p(\boldsymbol{t}_j \mid \boldsymbol{x}, \mathcal{P}, \boldsymbol{t}_{<j}) = \mathrm{LMHead}(\mathrm{LLMBackbone}([\mathbf{E}_v; \mathbf{E}_t; \mathrm{Embed}(\boldsymbol{t}_{<j})]))\,,$$

where $\mathrm{LMHead}(\cdot)$ maps the hidden state to a token distribution. The corresponding Next-Token Prediction (NTP) loss is

$$\mathcal{L}_{\mathrm{NTP}} = -\sum\nolimits_{j=1}^{T} \log p(\boldsymbol{t}_j \mid \boldsymbol{x}, \mathcal{P}, \boldsymbol{t}_{<j}).$$

3 The DualDistill Framework

Our goal is to transfer the generative recognition capability of an MLLM into flexible and efficient discriminative prediction without relying on human image-level annotations. To this end, we propose DualDistill, a progressive generation-to-discrimination framework that bridges prompt-conditioned generation and image-only discrimination through an intermediate prompt-aware discriminative pathway.

As illustrated in Figure 2, DualDistill consists of two complementary components. **Cross-Paradigm Distillation (CPD)** converts generative recognition into prompt-aware discriminative prediction while preserving the original generative learning signal of the MLLM. **Cross-Token Distillation (CTD)** transfers the resulting prompt-aware discriminative capability to an image-side classification token, enabling efficient prompt-free inference. We first introduce direct discriminative adaptation and analyze its limitations, which motivate this progressive design.

3.1 Direct Discriminative Adaptation

A natural way to improve the inference efficiency of an MLLM is to replace autoregressive label generation with direct discriminative prediction. Specifically, given an image $\boldsymbol{x}$, we append an image-side classification token $\langle\mathrm{CLS}_\mathcal{I}\rangle$ immediately after the visual tokens and use its hidden representation for multi-label classification. The input sequence is therefore $[\mathbf{E}_v; \mathbf{e}^\mathcal{I}_\mathrm{cls}]$, where $\mathbf{e}^\mathcal{I}_\mathrm{cls} = \mathrm{Embed}(\langle\mathrm{CLS}_\mathcal{I}\rangle)$ denotes the learnable embedding of the classification token. The corresponding hidden representation is obtained as $(\mathbf{H}_v, \mathbf{h}^\mathcal{I}_\mathrm{cls}) = \mathrm{LLMBackbone}\left([\mathbf{E}_v; \mathbf{e}^\mathcal{I}_\mathrm{cls}]\right)$, and the multi-label prediction is given by $\mathbf{p}^\mathcal{I} = \sigma\left(f^\mathcal{I}(\mathbf{h}^\mathcal{I}_\mathrm{cls})\right)$, where $f^\mathcal{I}$ denotes the classification head and $\sigma(\cdot)$ is the Sigmoid function. Under causal attention, $\langle\mathrm{CLS}_\mathcal{I}\rangle$ attends to all preceding visual tokens, and its resulting hidden state serves as a compact image-level representation. Importantly, this formulation requires neither the task-aware prompt nor autoregressive decoding during inference, making it substantially more efficient than generative recognition.

In our annotation-free setting, the image-side classifier is supervised using annotations generated by a frozen teacher MLLM. The generated textual responses are converted into multi-hot targets and

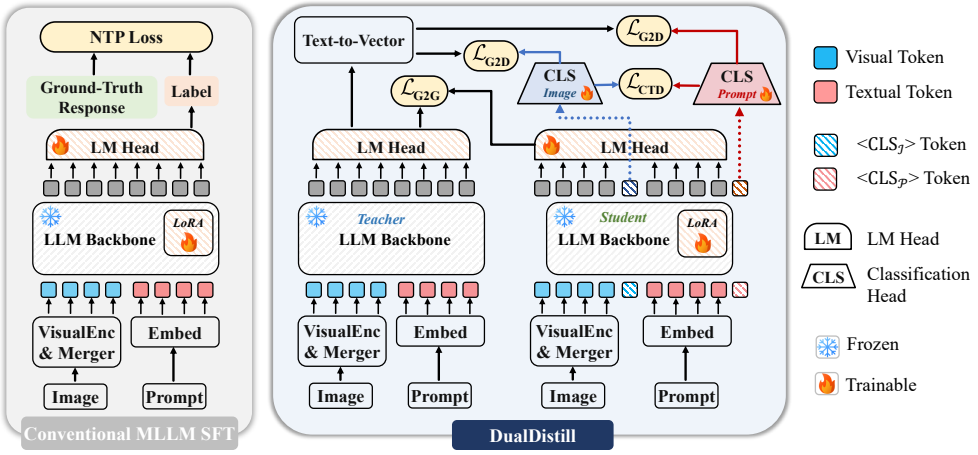


Figure 2: Conventional supervised MLLM fine-tuning versus DualDistill. DualDistill replaces human annotations with frozen-teacher responses, using CPD for discriminative adaptation and CTD for efficient prompt-free prediction.

used as discriminative supervision. The detailed formulation of this generative-to-discriminative transfer is introduced in Sec. 3.2. Despite its simplicity and inference efficiency, direct discriminative adaptation remains suboptimal for two reasons. First, unlike generative recognition conditioned on a task-aware prompt containing the candidate label vocabulary, the image-side classification token is placed before the prompt and therefore cannot attend to the subsequent prompt tokens under causal attention. Consequently, its hidden representation is learned without access to the task-specific semantic context provided by the candidate labels, which may limit its discriminative capability. Second, replacing autoregressive next-token prediction with a discriminative classification objective (see Eq. 1) introduces an optimization mismatch with the original learning paradigm of the MLLM, potentially limiting the exploitation of its pretrained knowledge.

3.2 Cross-Paradigm Distillation

The first step of DualDistill is to progressively convert prompt-conditioned generation into discriminative prediction through an intermediate prompt-aware pathway. To this end, we introduce a prompt-side classification token $\langle \mathrm{CLS}_\mathcal{P} \rangle$ after the task-aware prompt. Under causal attention, this token can access both the visual representations and the task-aware textual context, allowing it to retain the candidate-label semantics available during generative recognition while performing direct discriminative prediction. Based on this prompt-aware representation, we propose **Cross-Paradigm Distillation (CPD)**, which combines discriminative knowledge transfer with generative learning preservation. Specifically, the same teacher-generated response is exploited in two complementary forms: its multi-hot form provides discriminative supervision through Generative-to-Discriminative (G2D) distillation, while its original token sequence preserves autoregressive learning through Generative-to-Generative (G2G) distillation.

Generative-to-Discriminative Distillation We consider a frozen teacher MLLM $\mathcal{M}^\mathcal{T}$ and a student MLLM $\mathcal{M}^\mathcal{S}$. For each training image $\boldsymbol{x}$ and task-aware prompt $\mathcal{P}$, the teacher generates a textual annotation $\mathcal{M}^\mathcal{T}(\boldsymbol{x}, \mathcal{P})$, which typically takes the form of a comma-separated category list, such as "person, dog, cat". Given the category vocabulary $\mathcal{C} = \{C_k\}_{k=1}^q$, we define a Text-to-Vector (T2V) function $\phi(\cdot)$ that converts the generated annotation into a multi-hot vector $\hat{\boldsymbol{y}}^\mathcal{T} = \phi\big(\mathcal{M}^\mathcal{T}(\boldsymbol{x}, \mathcal{P})\big) \in \{0,1\}^q$. Specifically, $\hat{y}_k^\mathcal{T} = 1$ if category C_k appears in the teacher response and $\hat{y}_k^\mathcal{T} = 0$ otherwise. This converts the teacher's generative prediction into discriminative supervision without requiring any human annotations.

Given the dual-token input sequence, the student simultaneously obtains the hidden representations of the image-side and prompt-side classification tokens:

$$\left(\mathbf{H}_v, \mathbf{h}_{\mathrm{cls}}^\mathcal{I}, \mathbf{H}_t, \mathbf{h}_{\mathrm{cls}}^\mathcal{P}\right) = \mathrm{LLMBackbone}^\mathcal{S}\left([\mathbf{E}_v; \mathbf{e}_{\mathrm{cls}}^\mathcal{I}; \mathbf{E}_t; \mathbf{e}_{\mathrm{cls}}^\mathcal{P}]\right).$$

Under causal attention, the image-side token only accesses the preceding visual representations, whereas the prompt-side token can additionally exploit the task-aware textual context and thus serves as the intermediate prompt-aware representation in our progressive formulation. During training, we jointly maintain both pathways and supervise the two classification tokens with the teacher-generated multi-hot target.

The corresponding multi-label predictions are $\mathbf{p}^{\mathcal{I}} = \sigma\left(f^{\mathcal{I}}(\mathbf{h}^{\mathcal{I}}_{\text{cls}})\right)$ and $\mathbf{p}^{\mathcal{P}} = \sigma\left(f^{\mathcal{P}}(\mathbf{h}^{\mathcal{P}}_{\text{cls}})\right)$, where $f^{\mathcal{I}}$ and $f^{\mathcal{P}}$ denote the classification heads and $\sigma(\cdot)$ is the Sigmoid function. For $b \in \{\mathcal{I}, \mathcal{P}\}$, the Generative-to-Discriminative loss is defined as

$$\mathcal{L}^{b}_{\text{G2D}} = \sum\nolimits_{k=1}^{q} \hat{y}^{\mathcal{T}}_k \ell_1(p^b_k) + (1 - \hat{y}^{\mathcal{T}}_k)\ell_0(p^b_k), \tag{1}$$

where $\ell_1(\cdot)$ and $\ell_0(\cdot)$ denote the positive and negative terms of the asymmetric loss (ASL) (Ridnik et al., 2021) used for multi-label classification. The overall G2D objective is

$$\mathcal{L}_{\text{G2D}} = \mathcal{L}^{\mathcal{I}}_{\text{G2D}} + \mathcal{L}^{\mathcal{P}}_{\text{G2D}}.$$

Through G2D, the generative output of the teacher is converted from textual predictions into explicit discriminative supervision for both classification pathways.

Preserving the Generative Learning Paradigm G2D converts the teacher response into a discriminative target, but this conversion discards its original sequential form and optimizes the student solely through a classification objective. Such an objective differs substantially from the next-token prediction paradigm used during MLLM pretraining and instruction tuning. To complement the discriminative transfer, we therefore reuse the same teacher-generated response in its original textual form and introduce a Generative-to-Generative (G2G) objective.

Specifically, the teacher-generated annotation is tokenized using the tokenizer, resulting in a target sequence $\{\hat{\boldsymbol{t}}^{\mathcal{T}}_1, \ldots, \hat{\boldsymbol{t}}^{\mathcal{T}}_T\}$. During training, this target response is appended after the prompt-side classification token. The student then autoregressively predicts the teacher-generated response, and the G2G loss is defined as

$$\mathcal{L}_{\text{G2G}} = -\sum\nolimits_{j=1}^{T} \log p\left(\hat{\boldsymbol{t}}^{\mathcal{T}}_j \mid \boldsymbol{x}, \langle \text{CLS}_{\mathcal{I}} \rangle, \mathcal{P}, \langle \text{CLS}_{\mathcal{P}} \rangle, \hat{\boldsymbol{t}}^{\mathcal{T}}_{<j}\right).$$

Unlike G2D, which directly supervises the discriminative outputs, G2G retains the autoregressive learning signal of the pretrained MLLM. Since the generated response follows both classification tokens in the causal sequence, the generative objective also provides learning signals to the classification-token representations and the shared MLLM parameters. Together, G2D converts teacher-generated responses into explicit discriminative supervision, while G2G preserves their sequential form to retain the generative learning signal of the pretrained MLLM.

The overall CPD objective is therefore

$$\mathcal{L}_{\text{CPD}} = \mathcal{L}_{\text{G2D}} + \lambda \mathcal{L}_{\text{G2G}},$$

where λ controls the contribution of the generative objective. We set $\lambda = 1$ by default in all experiments unless otherwise specified.

3.3 Cross-Token Distillation

Although CPD establishes a stronger prompt-aware discriminative pathway, this pathway still requires encoding the task-aware prompt during inference. The second component of DualDistill therefore aims to transfer its prompt-aware discriminative capability to the image-side token, which retains the efficient prompt-free inference pathway introduced in Sec. 3.1. To this end, we propose **Cross-Token Distillation (CTD)**. Because the two classification tokens have asymmetric contextual access, the prompt-side token $\langle \text{CLS}_{\mathcal{P}} \rangle$ incorporates task-aware textual information that is unavailable to the image-side token $\langle \text{CLS}_{\mathcal{I}} \rangle$. We therefore treat the prediction of $\langle \text{CLS}_{\mathcal{P}} \rangle$ as a stop-gradient distillation target and transfer its prompt-aware discriminative knowledge to $\langle \text{CLS}_{\mathcal{I}} \rangle$.

For each category, the prediction $\mathbf{p}^{\mathcal{I}}_k$ and $\mathbf{p}^{\mathcal{P}}_k$ parameterize two Bernoulli distributions. Following Yang et al. (2023), we minimize the Kullback-Leibler (KL) divergence between the prompt-side and image-side predictions:

$$\mathcal{L}_{\text{CTD}} = \sum\nolimits_{k=1}^{q} D_{\text{KL}}\left(\text{sg}([p^{\mathcal{P}}_k, 1 - p^{\mathcal{P}}_k]) \parallel [p^{\mathcal{I}}_k, 1 - p^{\mathcal{I}}_k]\right),$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operation. Through this objective, the image-side token is encouraged to approximate the task-aware predictions of its prompt-side counterpart while preserving its image-only inference pathway. Importantly, both tokens are also directly supervised by the teacher-generated annotations through the G2D objective introduced in CPD. Therefore, CTD serves as an additional cross-token knowledge transfer mechanism rather than replacing direct discriminative supervision.

Combining CPD and CTD, the overall training objective of DualDistill is

$$\mathcal{L} = \mathcal{L}_{\mathrm{G2D}} + \lambda \mathcal{L}_{\mathrm{G2G}} + \beta \mathcal{L}_{\mathrm{CTD}},$$

where λ and β control the contributions of these two proposed objectives, respectively.

Theoretical interpretation of CTD. We further provide a multi-label perspective on how CTD transfers prompt-aware discriminative knowledge to the image-side pathway.

Proposition 1 (CTD-induced multi-label surrogate-risk bound) *Let $\mathbf{p}^{\mathcal{I}}, \mathbf{p}^{\mathcal{P}} \in [\epsilon, 1-\epsilon]^q$ denote the multi-label predictions of the image-side and prompt-side classifiers. For the CTD objective*

$$\mathcal{L}_{\mathrm{CTD}} = \sum\nolimits_{k=1}^{q} D_{\mathrm{KL}}\left(\mathrm{Bern}(p_k^{\mathcal{P}}) \| \mathrm{Bern}(p_k^{\mathcal{I}})\right),$$

the discrepancy between the two multi-label prediction vectors satisfies $\left\|\mathbf{p}^{\mathcal{I}} - \mathbf{p}^{\mathcal{P}}\right\|_2^2 \le \frac{1}{2}\mathcal{L}_{\mathrm{CTD}}$. Moreover, suppose the label-wise surrogate loss $\ell(y,p)$ is L_ℓ-Lipschitz on $[\epsilon, 1-\epsilon]$ with respect to p. Define the sample-wise multi-label risk of pathway $b \in \{\mathcal{I}, \mathcal{P}\}$ as $\mathcal{R}^b = \frac{1}{q}\sum_{k=1}^{q} \ell(y_k, p_k^b)$. Then,

$$\mathcal{R}^{\mathcal{I}} \le \mathcal{R}^{\mathcal{P}} + L_\ell \sqrt{\frac{\mathcal{L}_{\mathrm{CTD}}}{2q}}.$$

Proposition 1 shows that CTD explicitly controls the discrepancy between the entire multi-label prediction vectors of the two pathways. When the prompt-side classifier provides stronger task-aware predictions, minimizing $\mathcal{L}_{\mathrm{CTD}}$ bounds the excess surrogate risk of the image-side pathway relative to the prompt-side pathway. Therefore, CTD provides a principled mechanism for transferring the discriminative capability enabled by the task-aware prompt to the efficient image-only pathway.

4 Experiments

In this section, we first present the main experimental results and then provide further analysis.

4.1 Experimental Settings

Datasets We conduct experiments on three widely used multi-label recognition benchmarks: MS-COCO 2014 (Lin et al., 2014), Visual Genome (Krishna et al., 2017), and Objects365 (Shao et al., 2019). Following previous work (Chen et al., 2022; Xie et al., 2024), we preprocess Visual Genome into VG256 with 256 object categories. Detailed information is provided in the Appendix B.1.

Evaluation Metrics We evaluate both recognition performance and inference efficiency. For recognition, we report mean average precision (mAP), together with micro-F1, macro-F1, and sample-F1. For efficiency, we report throughput measured by images per second (IPS). Detailed metric definitions and inference settings are provided in the Appendix B.2.

Comparing Methods We compare DualDistill with a generative baseline and representative annotation-free methods. The **Generative Recognizer** fine-tunes the student MLLM using teacher-generated label sequences and performs autoregressive generation at inference time. We report the performance of the **Teacher MLLM** for reference. The compared annotation-free methods include **CDUL** (Abdelfattah et al., 2023) and **CCD** (Kim & Shim, 2025), as well as training-free methods **TagCLIP** (Lin et al., 2024), **DualPrompt** (Xie et al., 2026a), and **TagLLM** (Xie et al., 2026b).

Implementation We use Qwen2.5-VL-3B-Instruct and Qwen3-VL-2B-Instruct as the student backbones, and Qwen3-VL-8B-Instruct as the teacher MLLM. We fine-tune the LLM with LoRA together with the embeddings of $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$ and $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$ and the classification head. The learning rate is set to 0.001, and the CTD weight is fixed to $\beta = 0.01$ for all datasets. Additional implementation details are provided in the Appendix B.4.

Table 1: Comparison of different prediction configurations on COCO 2014, VG256, and Objects365. IPS denotes images processed per second. The best results are highlighted in **bold**.

COCO 2014						
Method	Model	IPS ↑	mAP ↑	Micro-F1 ↑	Macro-F1 ↑	Sample-F1 ↑
Generative Recognizer		15.70	–	85.46	85.29	87.75
DualDistill $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$	Qwen3-VL-2B	**45.65**	91.15	87.73	**86.86**	89.49
DualDistill $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$		30.66	**91.38**	**87.86**	86.48	**89.58**
Generative Recognizer		10.36	–	84.37	83.94	86.83
DualDistill $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$	Qwen2.5-VL-3B	**21.34**	89.13	86.21	84.79	88.16
DualDistill $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$		16.53	**89.33**	**86.45**	**85.34**	**88.34**
Teacher MLLM	Qwen3-VL-8B	–	–	85.26	84.52	87.40
VG256						
Method	Model	IPS ↑	mAP ↑	Micro-F1 ↑	Macro-F1 ↑	Sample-F1 ↑
Generative Recognizer		6.18	–	51.47	51.53	52.89
DualDistill $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$	Qwen3-VL-2B	**61.35**	51.72	64.29	53.90	63.71
DualDistill $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$		24.66	**52.12**	**64.44**	**54.25**	**63.79**
Generative Recognizer		5.08	–	50.58	50.44	52.04
DualDistill $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$	Qwen2.5-VL-3B	**29.63**	50.01	62.91	52.36	62.35
DualDistill $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$		15.03	**50.95**	**63.51**	**53.34**	**62.89**
Teacher MLLM	Qwen3-VL-8B	–	–	51.75	50.95	53.25
Objects365						
Method	Model	IPS ↑	mAP ↑	Micro-F1 ↑	Macro-F1 ↑	Sample-F1 ↑
Generative Recognizer		2.52	–	67.43	63.09	68.42
DualDistill $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$	Qwen3-VL-2B	**34.96**	64.98	69.30	63.10	70.87
DualDistill $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$		15.36	**65.77**	**69.49**	**63.78**	**71.01**
Generative Recognizer		2.13	–	65.15	60.32	66.30
DualDistill $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$	Qwen2.5-VL-3B	**16.84**	60.42	66.64	59.09	68.12
DualDistill $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$		9.38	**62.92**	**67.55**	**61.18**	**69.06**
Teacher MLLM	Qwen3-VL-8B	–	–	66.60	61.86	67.30

4.2 Comparison with MLLM Fine-Tuning Methods

Table 1 compares DualDistill with the Generative Recognizer on COCO 2014, VG256, and Objects365 in terms of inference efficiency and classification performance. Overall, DualDistill achieves substantially higher inference efficiency while maintaining strong recognition performance across different backbones and datasets. With Qwen3-VL-2B, the image-side token $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$ achieves approximately $2.9\times$, $9.9\times$, and $13.9\times$ higher throughput than the Generative Recognizer on COCO 2014, VG256, and Objects365, respectively, while the prompt-side token $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$ also provides clear speedups over autoregressive generation. In terms of recognition performance, both discriminative pathways consistently achieve higher Micro-F1 and Sample-F1 scores than the Generative Recognizer. Moreover, $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$ generally provides stronger classification performance by exploiting task-aware textual context, whereas $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$ offers a more favorable efficiency-performance trade-off through prompt-free inference. Notably, despite using substantially smaller student backbones, DualDistill can even outperform the Qwen3-VL-8B teacher MLLM on several metrics, demonstrating the effectiveness of transferring generative recognition capability into efficient discriminative prediction.

4.3 Comparison with the State-of-the-Art Methods

Table 2 compares DualDistill with existing annotation-free multi-label recognition methods on COCO 2014 and VG256. We consider both training-based annotation-free methods, including CDUL and CCD, and competitive training-free methods, including CLIP, TagCLIP, DualPrompt, and TagLLM. As shown in the table, DualDistill consistently achieves the best performance across

Table 2: Comparison with existing annotation-free multi-label recognition methods on COCO 2014 and VG256. A checkmark denotes methods requiring task-specific training. The best and second-best results are shown in **bold** and underline, respectively.

Method	Train	COCO 2014			VG256		
		mAP	Micro-F1	Macro-F1	mAP	Micro-F1	Macro-F1
CDUL	✓	69.20	–	–	–	–	–
CCD	✓	70.30	55.19	59.64	48.97	43.62	45.23
CLIP		64.92	61.48	58.66	33.67	43.93	35.78
TagCLIP		68.71	67.74	65.16	39.76	53.32	41.92
DualPrompt		71.11	68.73	66.03	40.48	53.37	42.09
TagLLM		–	79.16	81.36	–	44.03	46.44
DualDistill $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$	✓	<u>91.15</u>	<u>87.73</u>	**86.86**	<u>51.72</u>	<u>64.29</u>	<u>53.90</u>
DualDistill $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$	✓	**91.38**	**87.86**	<u>86.48</u>	**52.12**	**64.44**	**54.25**

Table 3: Ablation study of G2D, G2G, and CTD on VG256 and Objects365 using Qwen3-VL-2B. Best results are highlighted in **bold**.

G2D	G2G	CTD	**VG256**				**Objects365**			
			$\mathrm{CLS}_{\mathcal{I}}$		$\mathrm{CLS}_{\mathcal{P}}$		$\mathrm{CLS}_{\mathcal{I}}$		$\mathrm{CLS}_{\mathcal{P}}$	
			mAP	Macro-F1	mAP	Macro-F1	mAP	Macro-F1	mAP	Macro-F1
✓			46.50	47.56	50.27	52.15	62.07	58.45	62.42	59.13
✓	✓		48.40	50.59	51.14	53.22	61.19	57.02	64.91	62.58
✓	✓	✓	**51.72**	**53.90**	**52.12**	**54.25**	**64.98**	**63.10**	**65.77**	**63.78**

all reported metrics on both datasets, substantially outperforming existing methods. In particular, on COCO 2014, $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$ achieves 91.38% mAP, 87.86% Micro-F1, and 86.48% Macro-F1, considerably surpassing the strongest competing results. Similar improvements are observed on VG256, where $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$ reaches 52.12% mAP, 64.44% Micro-F1, and 54.25% Macro-F1. Notably, even the more efficient $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$ consistently outperforms all existing methods, demonstrating that DualDistill achieves strong recognition performance while retaining efficient discriminative inference.

4.4 Further Analysis

Ablation Studies To better understand the contribution of each component, Table 3 presents the ablation results of G2D, G2G, and CTD. G2G consistently improves the prompt-side pathway, while its effect on the image-side pathway is inconsistent. Further introducing CTD substantially improves $\langle CLS_{\mathcal{I}} \rangle$ and narrows the gap to $\langle CLS_{\mathcal{P}} \rangle$, validating the progressive transfer from prompt-aware to prompt-free discrimination. These results validate the effectiveness of progressively transferring prompt-aware discriminative capability to the more efficient image-side representation.

Effect of Candidate-Label Context As illustrated in Fig. 3, the prompt-side classifier produces predictions that better align with the ground-truth labels than the image-side classifier. Its strong attention to the candidate context and relevant label tokens further indicates that candidate-label semantics provide useful task-specific guidance that is unavailable to the image-side token.

Visualization Study on CTD We further visualize how CTD affects the relationship between $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$ and $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$. As shown in Figure 4, CTD generally increases their hidden-state similarity across layers, while the final-layer similarity distributions are clearly shifted toward higher values. These results demonstrate that CTD effectively transfers the prompt-aware discriminative knowledge of $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$ to the more efficient $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$.

Overall, these analyses validate the progressive design of DualDistill: candidate-label context strengthens the prompt-side pathway, while CTD effectively transfers this advantage to the efficient image-side pathway.

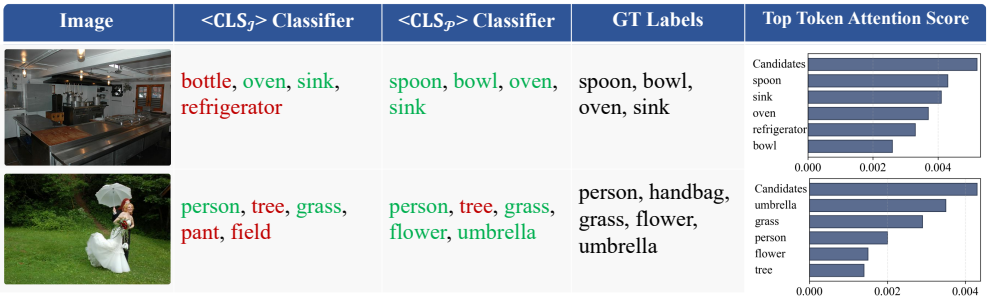
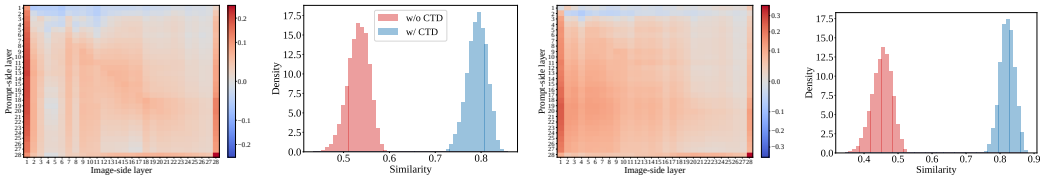


Figure 3: Qualitative comparison of image- and prompt-side classifiers without using CTD, highlighting the benefit of candidate-label context for prompt-side prediction.



(a) VG256 (b) VG256 (c) Objects365 (d) Objects365

Figure 4: Effect of CTD on the similarity between $\langle \mathrm{CLS}_\mathcal{P} \rangle$ and $\langle \mathrm{CLS}_\mathcal{I} \rangle$. (a,c) Layer-wise similarity gains induced by CTD on VG256 and Objects365. (b,d) Final-layer similarity distributions, showing that CTD consistently improves cross-token alignment.

5 Related Works

Annotation-free multi-label visual recognition mainly follows two research directions: training-based and training-free methods. Training-based methods first exploit the strong zero-shot recognition capability of CLIP to generate pseudo labels, and then train a downstream classifier based on these pseudo labels (Abdelfattah et al., 2023; Kim & Shim, 2025). Training-free methods focus only on the first step, seeking to better exploit CLIP's zero-shot recognition capability through enhanced feature learning (Lin et al., 2024) or improved prompting strategies (Xie et al., 2026a), without training a downstream classifier. More recently, MLLMs have been leveraged as stronger annotators (Xie et al., 2026b), leading to significant performance improvements.

MLLM fine-tuning has become a standard paradigm for adapting general-purpose vision-language models to downstream tasks. Existing studies investigate alleviating pre-trained knowledge forgetting during downstream adaptation (Zhai et al., 2024; Huang et al., 2025), as well as efficient fine-tuning strategies for different downstream tasks, such as fine-grained classification (He et al., 2025), temporal action understanding (Lin & Shou, 2025; Pang et al., 2026), and emotion understanding (Lian et al., 2025). In contrast, we focus on transferring the generative recognition capability of MLLMs into efficient discriminative prediction under annotation-free supervision.

6 Conclusion

We present DualDistill, a dual-distillation framework for inference-efficient and annotation-free multi-label classification in MLLMs. DualDistill introduces CPD to transfer the generative annotation capability of a teacher MLLM into a classification head built on the prompt-side classification token, enabling efficient one-forward prediction without human annotation. It then employs CTD to distill knowledge from the prompt-side token into an image-side token, thereby further improving inference efficiency while maintaining strong classification performance. We also incorporate a generative-to-generative objective to better optimize MLLM-related parameters. Extensive experiments show that DualDistill achieves a superior balance between inference efficiency and classification performance, substantially outperforming strong baselines and state-of-the-art methods.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing and to assist with retrieval and discovery of relevant literature. We did not use generative AI tools for other tasks requiring disclosure. All AI-assisted text was reviewed and revised by the authors, and suggested literature was manually verified against the original sources. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on annotation-free multi-label visual recognition using existing benchmark datasets and does not involve human-subject studies or the collection of new personal data. We do not release any new dataset containing sensitive information. We are not aware of any significant ethical concerns arising from the methodology or experiments presented in this work.

REPRODUCIBILITY STATEMENT

We provide the main architectural and training details of DualDistill in Secs. 3 and 4, including the model formulation, optimization objectives, datasets, data splits, and implementation settings. Additional experimental details, hyperparameters, and ablation settings are provided in the appendix and supplementary materials. Anonymous source code and configuration files are also provided in the supplementary materials to facilitate reproduction of the reported results.

REFERENCES

Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1348–1357, October 2023.

Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, and Liang Lin. Structured semantic transfer for multi-label recognition with partial labels. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:339–346, 2022.

Ruiqi Du, Xu Tang, Jingjing Ma, Xiangrong Zhang, and Licheng Jiao. Mlmamba: A mamba-based efficient network for multi-label remote sensing scene classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7):6245–6258, 2025.

Beatriz Quintino Ferreira, Joao P. Costeira, and Joao P. Gomes. Explainable noisy label flipping for multi-label fashion image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3916–3920, 2021.

Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=p3NKpom1VL.

Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. Learn from downstream and be yourself in multimodal large language models fine-tuning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=FKqmIAnkrb.

Naoto Inoue, Edgar Simo-Serra, Toshihiko Yamasaki, and Hiroshi Ishikawa. Multi-label fashion image classification with minimal human supervision. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017.

Dongseob Kim and Hyunjung Shim. Classifier-guided clip distillation for unsupervised multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4661–4671, June 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 36993–37014, 2025.

Kevin Qinghong Lin and Mike Zheng Shou. Vlog: Video-language models by generative retrieval of narration vocabulary. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 3218–3228, June 2025.

Lequan Lin, Dai Shi, Andi Han, Feng Chen, Qiuzheng Chen, Jiawen Li, Zhaoyang Li, Jiyuan Zhang, Zhenbang Sun, and Junbin Gao. ACT as human: Multimodal large language model data annotation with critical thinking. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, 2014.

Yuqi Lin, Minghao Chen, Kaipeng Zhang, Hengjia Li, Mingming Li, Zheng Yang, Dongqin Lv, Binbin Lin, Haifeng Liu, and Deng Cai. Tagclip: a local-to-global framework to enhance open-vocabulary multi-label classification of clip without training. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, 2024.

Zhanzhong Pang, Dibyadip Chatterjee, Fadime Sener, and Angela Yao. On discriminative vs. generative classifiers: Rethinking MLLMs for action understanding. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ppceQOZrAX.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 82–91, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 930–957, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024.

Ming-Kun Xie, Jia-Hao Xiao, Pei Peng, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Counterfactual reasoning for multi-label image classification via patching-based training. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=lQIN9ZyMLz.

Ming-Kun Xie, Zhiqiang Kou, Zhongnian Li, Gang Niu, and Masashi Sugiyama. Unlocking the power of co-occurrence in CLIP: A dualprompt-driven method for training-free zero-shot multi-label classification. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Ming-Kun Xie, Jia-Hao Xiao, Zhiqiang Kou, Zhongnian Li, Gang Niu, and Masashi Sugiyama. Are multimodal large language models good annotators for image tagging?, 2026b. URL `https://arxiv.org/abs/2602.20972`.

Penghui Yang, Ming-Kun Xie, Chen-Chen Zong, Lei Feng, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Multi-label knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17271–17280, 2023.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, volume 234, pp. 202–227, 2024.

Yuhan Zhang, Luyang Luo, Qi Dou, and Pheng-Ann Heng. Triplet attention and dual-pool contrastive learning for clinic-driven multi-label medical image classification. *Medical Image Analysis*, 86:102772, 2023.

A Proof of Proposition 1

For each category k, the prompt-side and image-side predictions define two Bernoulli distributions, $\mathrm{Bern}(p_k^P)$ and $\mathrm{Bern}(p_k^I)$. By Pinsker's inequality,

$$\left|p_k^I - p_k^P\right| \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}\left(\mathrm{Bern}(p_k^P) \| \mathrm{Bern}(p_k^I)\right)}.$$

Squaring both sides gives

$$\left(p_k^I - p_k^P\right)^2 \le \frac{1}{2} D_{\mathrm{KL}}\left(\mathrm{Bern}(p_k^P) \| \mathrm{Bern}(p_k^I)\right).$$

Summing over all q categories yields

$$\begin{aligned} \left\|\mathbf{p}^I - \mathbf{p}^P\right\|_2^2 &= \sum_{k=1}^{q} \left(p_k^I - p_k^P\right)^2 \\ &\le \frac{1}{2} \sum_{k=1}^{q} D_{\mathrm{KL}}\left(\mathrm{Bern}(p_k^P) \| \mathrm{Bern}(p_k^I)\right) \\ &= \frac{1}{2} \mathcal{L}_{\mathrm{CTD}}. \end{aligned} \tag{2}$$

We next consider the multi-label surrogate risk. Since $\ell(y, p)$ is L_ℓ-Lipschitz with respect to p,

$$\begin{aligned} \left|\mathcal{R}^I - \mathcal{R}^P\right| &\le \frac{L_\ell}{q} \sum_{k=1}^{q} \left|p_k^I - p_k^P\right| \\ &= \frac{L_\ell}{q} \left\|\mathbf{p}^I - \mathbf{p}^P\right\|_1. \end{aligned}$$

By the Cauchy–Schwarz inequality,

$$\left\|\mathbf{p}^I - \mathbf{p}^P\right\|_1 \le \sqrt{q} \left\|\mathbf{p}^I - \mathbf{p}^P\right\|_2.$$

Combining this result with Eq. 2, we obtain

$$\left|\mathcal{R}^I - \mathcal{R}^P\right| \le L_\ell \sqrt{\frac{\mathcal{L}_{\mathrm{CTD}}}{2q}},$$

which immediately implies

$$\mathcal{R}^I \le \mathcal{R}^P + L_\ell \sqrt{\frac{\mathcal{L}_{\mathrm{CTD}}}{2q}}.$$

This completes the proof.

B Additional Experimental Details

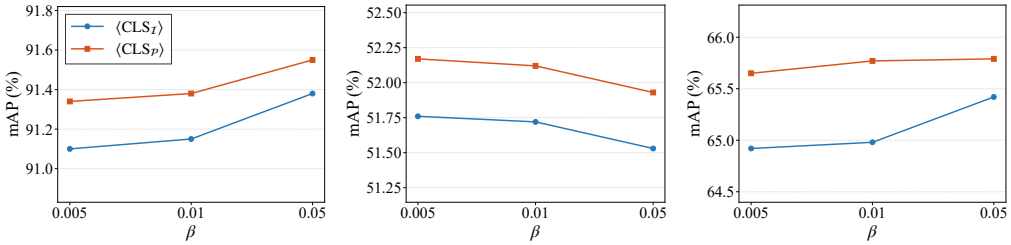
B.1 Datasets

We evaluate DualDistill on three benchmark datasets: MS-COCO 2014 (Lin et al., 2014), Visual Genome (Krishna et al., 2017), and Objects365 (Shao et al., 2019).

MS-COCO 2014 MS-COCO 2014 contains 80 object categories, with 82,081 images in the training set and 40,137 images in the validation set. Each image contains an average of approximately 2.9 object labels.

Visual Genome The original Visual Genome dataset contains 108,249 images and 80,138 object categories. Following the preprocessing protocol adopted in previous multi-label recognition studies (Chen et al., 2022; Xie et al., 2024), we construct VG256 containing 106,702 images and 256 object categories. We randomly split the dataset into a 70% training set with 74,691 images and a 30% validation set with 32,011 images.

Objects365 Objects365 contains 365 object categories. In our experiments, we only use the official validation set, which contains approximately 80,000 images, rather than the original training set. We randomly split these images into training and validation subsets with a 7:3 ratio.



(a) COCO 2014 (b) VG256 (c) Objects365

Figure 5: Sensitivity analysis of the CTD weight β on COCO 2014, VG256, and Objects365. Both the image-side token $\langle \mathrm{CLS}_\mathcal{I} \rangle$ and prompt-side token $\langle \mathrm{CLS}_\mathcal{P} \rangle$ remain stable across different values of β.

B.2 Evaluation Metrics

We evaluate the models from two perspectives: recognition performance and inference efficiency.

For score-based recognition, we compute the average precision (AP) for each category and report mean average precision (mAP) over all categories. For binary multi-label predictions, we additionally report micro-F1, macro-F1, and sample-F1 to evaluate recognition performance from complementary perspectives.

For inference efficiency, we use images per second (IPS) as the throughput metric, analogous to the frames-per-second (FPS) metric commonly used in video processing. IPS measures the number of images processed by the model per second. All inference speeds are measured on a single NVIDIA A100 80GB Tensor Core GPU using BF16 precision. We use a batch size of 32 for COCO 2014 and VG256, and 16 for Objects365. For the Generative Recognizer, we adopt greedy decoding without sampling and terminate generation when the end-of-sequence token is produced, with the maximum generation length set to 128 tokens. In contrast, the discriminative pathways directly produce multi-label predictions in a single forward pass without autoregressive decoding; the $\langle \mathrm{CLS}_\mathcal{I} \rangle$ pathway additionally removes task-prompt encoding, whereas $\langle \mathrm{CLS}_\mathcal{P} \rangle$ retains the same task-aware prompt used by the Generative Recognizer.

B.3 Compared Methods

We compare DualDistill against both generative and discriminative annotation-free recognition approaches.

Generative Recognizer We construct a generative baseline using the same student MLLM backbone as DualDistill. The model is fine-tuned with the next-token prediction (NTP) objective using label sequences generated by the teacher MLLM. During inference, it autoregressively generates the predicted label sequence through the LM head. The generated text is subsequently converted into a multi-hot prediction vector using the text-to-vector (T2V) function. This baseline represents the conventional generative recognition paradigm and enables a direct comparison with the discriminative inference introduced by DualDistill.

Teacher MLLM We additionally report the recognition performance of the frozen teacher MLLM, Qwen3-VL-8B-Instruct, as a reference for the quality of the generative supervision used to train the student models.

Annotation-free Methods We compare with CDUL (Abdelfattah et al., 2023) and CCD (Kim & Shim, 2025), which construct pseudo supervision without relying on human-annotated multi-label training labels. We further include representative training-free approaches, including TagCLIP (Lin et al., 2024), DualPrompt (Xie et al., 2026a), and TagLLM (Xie et al., 2026b).

B.4 Implementation Details

We adopt Qwen2.5-VL-3B-Instruct and Qwen3-VL-2B-Instruct as the student backbones for DualDistill, while Qwen3-VL-8B-Instruct serves as the frozen teacher MLLM for generating annotation-free supervision.

During training, the vision encoder and visual merger are kept frozen. We optimize the LLM using LoRA, together with the embeddings of the image-side token $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$, the prompt-side token $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$, and the classification heads. For LoRA, we use rank $r = 8$ and scaling factor $\alpha = 16$. The learning rate is set to 0.001 for all trainable parameters. The weight of the cross-token distillation objective is fixed to $\beta = 0.01$ across all datasets.

C Sensitivity Analysis of the CTD Weight

As shown in Fig. 5, both $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$ and $\langle \mathrm{CLS}_{\mathcal{P}} \rangle$ exhibit relatively stable performance over $\beta \in \{0.005, 0.01, 0.05\}$ on all three datasets, indicating that DualDistill is not highly sensitive to the exact choice of the CTD weight. On COCO 2014 and Objects365, increasing β slightly improves the image-side pathway, suggesting that stronger cross-token transfer can further benefit $\langle \mathrm{CLS}_{\mathcal{I}} \rangle$. In contrast, VG256 shows a mild performance decrease as β increases, implying that excessive distillation may slightly constrain the image-side pathway on this dataset. Despite these dataset-specific trends, the overall variation remains small for both classification tokens, demonstrating the robustness of DualDistill to β. We therefore use $\beta = 0.01$ as a common default across all datasets. These results suggest that CTD provides a stable transfer signal over a broad range of weights, while the preferred transfer strength may vary slightly with the dataset.