Causal Denoising Prototypical Network for Few-Shot Multi-label Aspect Category Detection

Jin Cui[†], **Xinfeng Wang**[†], **Fumiyo Fukumoto**[‡], and **Yoshimi Suzuki**[‡]
[†]Graduate School of Engineering
[‡]Interdisciplinary Graduate School
University of Yamanashi, Kofu, Japan
{g22dtsa5, g22dtsa7, ysuzuki, fukumoto}@yamanashi.ac.jp

Abstract

The multi-label aspect category detection (MACD) task has attracted great attention in sentiment analysis. Many recent methods have formulated the MACD task by learning robust prototypes to represent categories with limited support samples. However, few of them address the noise categories in the support set that hinder their models from effective prototype generations. To this end, we propose a causal denoising prototypical network (CDPN) for few-shot MACD. We reveal the underlying relation between causal inference and contrastive learning, and present causal contrastive learning (CCL) using discrete and continuous noise as negative samples. We empirically found that CCL can (1) prevent models from overly predicting more categories and (2) mitigate semantic ambiguity issues among categories. Experimental results show that CDPN outperforms competitive baselines. Our code is available online: https://github.com/cuijin-23/CPDN.

1 Introduction

With the rapid growth of online services, multi-label aspect category detection (MACD) that detects aspect categories within a given sentence assigned multiple categories (Li et al., 2020; Zhang et al., 2022b; Kamila et al., 2022), has gained considerable attention. Due to the success of few-shot learning (Wang et al., 2020; Song et al., 2023; Zhou et al., 2023; Nookala et al., 2023) and meta-learning (ML) (Zhang et al., 2022a; Liang et al., 2023), Hu et al. (2021) have formulated MACD task with a meta-task which is illustrated in Fig. 1. It aims to predict the class label of the query set by leveraging the support set. Thereafter, Zhao et al. (2022) and Liu et al. (2022) propose label-driven prototype denoising models. Zhao et al. (2023) employ sample-set operations based on ML for MACD tasks. Peng et al. (2024) utilize an ML-based framework with variational distribution

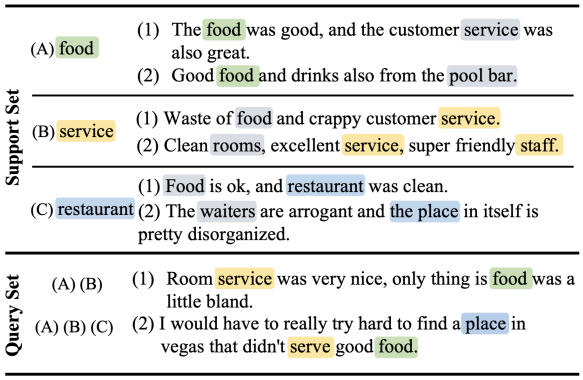


Figure 1: An example of a 3-way 2-shot meta-task. The colored boxes denote the target aspect categories, and the gray boxes refer to words that cause noisy categories.

inference, achieving state-of-the-art (SOTA) performance in few-shot MACD. However, the aforementioned approaches ignore the fact that sentences in the support set often mention multiple categories, and most of these are noise which deteriorates the overall performance. For instance, support samples of (B) in Fig. 1 mention "food" and "room", which are noise categories for learning the prototype of "service". It indicates that not only the target category (i.e., "service") but also the non-target categories (i.e., "food" and "room") in support samples affect the learning of the prototype.

The causal interference mechanism has been widely utilized to reduce the negative impact of intermediate factors (e.g., noise or spurious correlations) (Chen et al., 2021; Feng et al., 2021; Cao et al., 2022; Tian et al., 2022; Tu et al., 2023). As the casual graph shown in Fig. 2 (a), the treatment variable X directly affects the response variable Y and indirectly affects Y through $\mathcal{N}$, where $\mathcal{N}$ refers to the undesired factor. They mitigate the impact of $\mathcal{N}$ by using counterfactual intervention to force that $Y = y$ remains the same when changing $\mathcal{N}$ from η to η^* in (d), where y and η are realizations of Y and $\mathcal{N}$. However, few approaches utilize

Findings of the Association for Computational Linguistics: ACL 2025, pages 7091–7104

causal interference to denoise the representations of support samples for MACD tasks.

Several recent approaches have utilized contrastive learning (CL) (Zhao et al., 2022; Liu et al., 2022) to push prototypes of various categories away, achieving remarkable results. They regard the representations of other categories as negative pairs for contrast. However, they pay no attention to integrating causal inference and CL, although there exist underlying yet strong relations between them, in the sense that (1) the noise-free state of $\mathcal{N} = \eta_{\emptyset}$ can be regarded as a positive sample for CL; (2) the way to reinforce the cohesion between positive pairs and diversity between negative pairs in CL can be utilized to predict Y for causal inference (Wang et al., 2021; Cao et al., 2022).

Motivated by the observations, we propose a causal denoising prototypical network (CDPN) for few-shot MACD. Specifically, we integrate causal inference with CL to effectively denoise the representation of support samples. Different from related works that regard the representations of other categories as negative pairs (Zhao et al., 2022; Liu et al., 2022), our approach utilizes the noise representations extracted from support samples for causal contrastive learning (CCL).

Moreover, we introduce discrete and continuous noise as negative samples for CCL. The discrete noise is generated by a negative-label-driven attention mechanism, and the continuous noise is obtained by applying dropout at a substantial rate on the representations of support samples. We empirically found that discrete noise representation facilitates better convergence by incorporating multiple negative labels. Conversely, the advantage of continuous noise representation lies in its ability to be obtained without requiring negative labels.

The main contributions of our work can be summarized as follows: (1) We highlight the underlying yet strong relations between causal inference and contrastive learning for MACD tasks and thus propose a novel causal contrastive learning; (2) We explore discrete and continuous noise as negative samples for CCL to denoise the representations of support samples; This enables CCL to work well without negative labels and perform even better with them; (3) Extensive experiments show that CDPN outperforms SOTA baselines in the MACD task. We discuss its efficacy compared to some LLMs (i.e., Llama2-7B, Llama3-8B, GPT-3.5-turbo and GPT-4o-mini). We also found that these LLMs tend to overly infer more aspect categories, while CCL can prevent our CDPN from the issue.

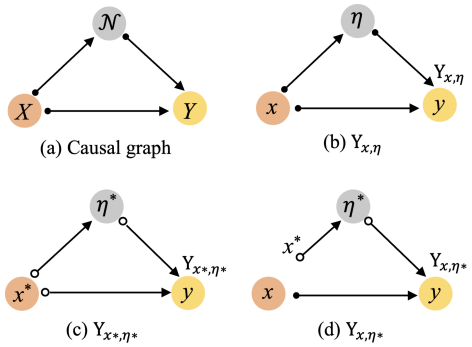


Figure 2: (a) The causal graph where X directly affects Y and indirectly affects Y through $\mathcal{N}$. (b) A causal graph with particular realizations. (c) A causal intervention $do(\mathcal{N} = \eta^*)$. (d) A counterfactual situation where $X = x$ remains the same while forcing $\mathcal{N} = \eta^*$.

2 Related Work

2.1 Few-Shot Learning

The few-shot learning has been widely explored to improve sentiment analysis (Hosseini-Asl et al., 2022; Wang et al., 2023; Liang et al., 2023). Recently, Hu et al. (2021) formulate MACD tasks with a prototypical network of meta-learning (ML) (Snell et al., 2017; Gao et al., 2019; Chen et al., 2023a). Thereafter, Zhao et al. (2022) and Liu et al. (2022) propose a label-driven prototype denoising model. Wang and Iwaihara (2023) present an ML-based dual-attention approach. Zhao et al. (2023) attempt to employ sample-set operations based on ML for MACD tasks. Peng et al. (2024) present an ML-based framework with variational distribution inference, achieving SOTA performance in MACD.

2.2 Contrastive Learning

Contrastive learning (CL) has been extensively explored for prototype learning, such as target-aware prototypical graph CL-based model (Liang et al., 2022), and ContrastNet (Chen et al., 2022) that reduce prototype contradictions among similar classes with CL. In MACD tasks, Liu et al. (2022) integrate contrastive learning into their prototypical network, and Zhao et al. (2022) propose label-weighted CL to highlight the semantic correlations among similar aspects.

2.3 Causal Inference

The counterfactual inference has become a prevalent solution to address undesirable bias and concerns in variable natural language understanding

(Tian et al., 2022; Udomcharoenchaikit et al., 2022), recommendation (Zhang et al., 2021; Li et al., 2024), text classification (Qian et al., 2021; Veitch et al., 2021; Wang and Culotta, 2021), and sentiment analysis (Wang et al., 2022; Yuan et al., 2022; Chang et al., 2024; Wu et al., 2024).

3 Preliminaries

3.1 Problem Formulation

We define the meta-task learning-based few-shot MACD task, following (Zhao et al., 2022)'s work. Suppose in an N-way K-shot meta-task, the support set is $\mathcal{X} = \{(\chi_1^i, ..., \chi_K^i), z^i\}_{i=1}^N$, where each χ_i denotes a sentence corresponding to the aspect category label z^i. Each query instance in the query set $\mathcal{Q}$ is defined as (χ_q, z_q), where z_q ($z_q \in \mathbb{R}^N$, $z_q^i \in \{0, 1\}$) is the query label indicating the aspects in χ_q out of N classes.

3.2 Counterfactual Inference

Causal Inference. Existing methods typically use a causal graph to capture the relationships between variables (Mitrovic et al., 2020; Tu et al., 2023). They exclude the impact of $\mathcal{N}$ by using a causal intervention to maintain $Y = y$ while changing $\mathcal{N}$ from η to η^*:

$$P(y|X = x, \mathcal{N} = \eta) = P(y|X = x, \mathcal{N} = \eta^*).$$

Specifically, they exert a causal intervention on $\mathcal{N}$ with a do-operator $do(\mathcal{N} = \eta^*)$ in Fig. 2 (c), which forcibly substitute η with η^* to obtain a counterfactual state $Y_{x,\eta^*} = f_Y(X = x, \mathcal{N} = \eta^*)$ in (d).

Causal Effect. The causal effect studies the causal influence among variables (Chen et al., 2021; Cao et al., 2022; Xu et al., 2023). The total effect (TE) measures the change of the response variable Y when the treatment variable X changes from x^* to x. The TE of $X = x$ on Y is denoted as:

$$\text{TE} = Y_{x,\eta} - Y_{x^*,\eta^*}.$$

The TE is often regarded as the sum of natural direct effect (NDE) and total indirect effect (TIE) (Wang et al., 2021; Chen et al., 2023b). The NDE of $X = x$ on Y refers to the change of Y when changing X from x^* to x and forcing $\mathcal{N} = \eta^*$:

$$\text{NDE} = Y_{x,\eta^*} - Y_{x^*,\eta^*}.$$

The TIE of $X = x$ on Y denotes the change of Y when only the $\mathcal{N}$ changes from η to $\eta*$, which can be obtained by subtracting NDE from TE:

$$\text{TIE} = \text{TE} - \text{NDE} = Y_{x,\eta} - Y_{x,\eta^*}.$$

As such, through causal intervention $do(\mathcal{N} = \eta^*)$, they can maximize the TIE to emphasize the impact of $\mathcal{N}$ (Tian et al., 2022) or minimize the TIE to mitigate the undesirable effects of factor $\mathcal{N}$ (Wang et al., 2021; Sun et al., 2022).

4 Methodology

Fig. 3 illustrates the framework of our CDPN, which consists of (i) prototypical representation learning and (ii) causal representation denoising.

4.1 Prototypical Representation Learning

We utilize support samples to generate prototypical representations of categories and estimate the distance between the query representation to identify aspect categories. Formally, given a sentence χ_i with T words, we define the embedding matrix as $H_i = [h_{i,1}, h_{i,2}, ..., h_{i,t}] \in \mathbb{R}^{d \times t}$, where d refers to the embedding size. We then obtain the representation s_i of the sentence, which is given by:

$$\begin{aligned} s_i &= \text{softmax}(a_i)H_i, \\ a_i &= \tanh(\bar{H}_i^\top (w_1 H_i + b_1)), \end{aligned} \tag{1}$$

where a_i refers to an self-attention score, $\bar{H}_i = \frac{1}{T}\sum_{t=1}^{T} h_{i,t}$ represents a query vector in self-attention, and w_1 and b_1 are trainable parameters.

Subsequently, in the N-way K-shot setting, we obtain the embedding $\mathcal{S} = \{s_1^1, s_2^1, ..., s_K^1, ..., s_K^N\}$ of the support samples $\mathcal{X}$. Similarly, the category embedding of support samples, i.e., the label text description information, can be represented as $C = \{c_1, c_2, ...c_N\}$. We calculate the prototype p^n through K instances for aspect category n by:

$$\begin{aligned} p^n &= \frac{1}{K}\sum_{i=1}^{K} \alpha^n s_i^n, \\ \alpha^n &= \text{softmax}(\cos(c^n, s_i^n)), \end{aligned} \tag{2}$$

where $\cos(c^n, s_i^n)$ refers to the cosine similarity between the sentence embedding s_i^n and category embedding c^n, $i \in [1, K]$ and $n \in [1, N]$.

Given query embedding $q \in \mathcal{Q}$, we compute the probability $P(z = z^n|q, \mathcal{S})$ to predict its aspect label via negative squared Euclidean distance.

$$P(z = z^n|q, \mathcal{S}) = \frac{\exp(-||q - p^n||_2^2)}{\sum_{j=1}^{N} \exp(-||q - p^j||_2^2)}. \tag{3}$$

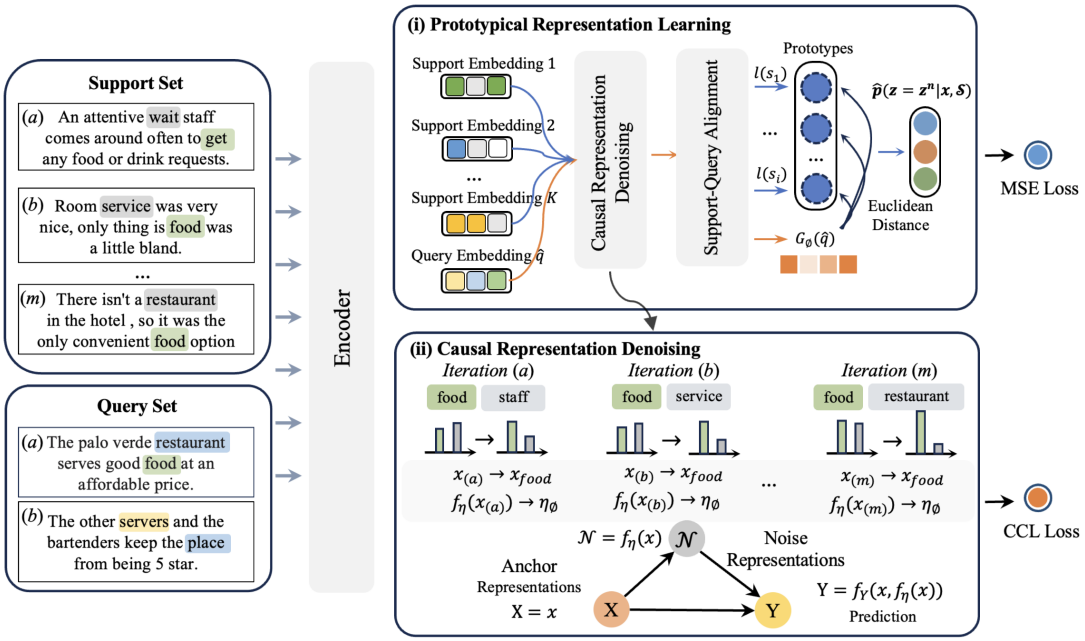


Figure 3: The architecture of our causal denoising prototypical network (CDPN).

We calculate the mean square error (MSE) loss on all query samples in $\mathcal{Q}$ to optimize CDPN:

$$\mathcal{L}_{acd} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \sum_{n=1}^{N} (P(z = z^n | q, \mathcal{S}) - \mathbf{z}^n)^2, \quad (4)$$

where $\mathbf{z}^n$ is the ground-truth label of category z^n.

4.2 Causal Representation Denoising

We introduce causal intervention to denoise representation of support sample s_i, which we call anchor representation. We define the TIE by:

$$\begin{aligned} \text{TIE} &= Y_{x,\eta} - Y_{x,\eta^*} \\ &= f_Y(x, f_\eta(x)) - f_Y(x, do(\mathcal{N} = \eta_\emptyset)) \\ &= f_Y(x, f_\eta(x)) - f_Y(x, \eta_\emptyset), \end{aligned} \quad (5)$$

where $do(\mathcal{N} = \eta_\emptyset)$ represents a causal intervention that forcefully assigns $\mathcal{N}$ to a reference status $\eta_\emptyset$ that is free of noise, resulting in a post-intervention prediction $f_Y(x, do(\mathcal{N} = \eta_\emptyset))$.

Based on the definition, we propose a causal contrastive learning approach for denoising. Specifically, we define X as the anchor representation, $\mathcal{N}$ as the representation with varying levels of noise extracted from X, and Y as the prediction with $f_Y(X = x, \mathcal{N} = \eta)$. Note that support samples for the target category (e.g., "food") include diverse noises (e.g., "staff," "service," and "restaurant,") while all samples share the core representation (e.g., "food"). Thus, we define $f_Y(x, \eta) = \text{sim}(x, \eta)/\mu$ to guarantee that the model consistently focuses on the feature of the target category (e.g., "food") when the $\mathcal{N}$ changes from $f_\eta(x)$ (e.g., "staff," "service," or "restaurant") to $\eta_\emptyset$, making $\eta_\emptyset$ converge to the core representation of x in Fig. 3. Here, μ refers to the temperature and $\text{sim}(\cdot)$ refers to cosine similarity. Further, we regard the expected noise-free representation $\eta_\emptyset$ as a positive sample $s_i^{(+)}$ and $f_\eta(x)$ as a negative sample that is extracted from x. The vanilla contrastive learning is given by:

$$\begin{aligned} \mathcal{L}_{ccl} &= \mathbb{E}_{s_i \sim \mathcal{S}} \left[-\log \frac{\exp(\text{sim}(s_i, s_i^{(+)})/\mu)}{\sum_{j=1}^{J} \exp(\text{sim}(s_i, s_{i,j}^{(-)})/\mu))} \right] \\ &= \underbrace{\mathbb{E}_{s_i \sim \mathcal{S}} \left[\log \sum_{j=1}^{J} \exp(\text{sim}(s_i, s_{i,j}^{(-)})/\mu)) \right]}_{Y_{x,\eta} \neq f_Y(x, f_\eta(x))} \\ &\quad - \underbrace{\mathbb{E}_{s_i \sim \mathcal{S}} \left[\text{sim}(s_i, s_i^{(+)})/\mu \right]}_{Y_{x,\eta^*} = f_Y(x, \eta_\emptyset)}. \end{aligned} \quad (6)$$

Through Eq. (6), we observed two issues: (1) $Y_{x,\eta}$ does not conform to $f_Y(x, f_\eta(x))$ and (2) $Y_{x,\eta}$ is affected by the number of J negative samples, which contradicts the definition of causal effect, i.e., $Y_{x,\eta}$ and Y_{x,η^*} should correspond to two distinct states of Y caused by $f_\eta(x)$ and $\eta_\emptyset$. To this end, we provide novel definitions of positive and negative samples. Firstly, we use a label-driven attention technique to obtain positive samples as follows:

$$\begin{aligned} s_i^{(+)} &= \beta_n^\top s_i, \\ \beta_n &= \text{softmax}(\tanh(w_2\, c_n) w_3)), \end{aligned} \quad (7)$$

where w_2 and w_3 denote weight matrices and $c_n \in \mathbb{R}^{d' \times d}$ is the embedding of the label word (e.g., "food_drink" is often separated by tokenizers as "food,""_," and "drink"), where d' denotes the padded dimension size.

Secondly, we generate discrete and continuous noise representations as negative samples. For the discrete noise representations, we merge the negative labels in the support set (e.g., "service," "restaurant," and "price") into a single text (e.g., "service [SEP] restaurant [SEP] price"), the representation of which is denoted by $c'_n \in \mathbb{R}^{d' \times d}$. We obtain the representation of negative samples by:

$$\begin{aligned} s_i^{(-)} &= {\beta'_n}^{\top} s_i, \\ \beta'_n &= \text{softmax}(\tanh(w_4\, c'_n) w_5), \end{aligned} \tag{8}$$

where w_4 and w_5 are weight matrices. For continuous noise representations, following (Wang et al., 2024), we utilize dropout at rate r on the anchor representation, which is given by:

$$s_i^{(-)} = \text{dropout}(s_i, r). \tag{9}$$

As each of the discrete and continuous representations corresponds to a negative sample (i.e., J=1), the causal contrastive learning can be rewritten as:

$$\begin{aligned} \mathcal{L}_{ccl} =& \mathbb{E}_{s_i \sim \mathcal{S}} \left[-\log \frac{\exp(\text{sim}(s_i, s_i^{(+)})/\mu)}{\sum_{j=1}^{J} \exp(\text{sim}(s_i, s_i^{(-)})/\mu)} \right] \\ =& \mathbb{E}_{s_i \sim \mathcal{S}} \underbrace{\left[\text{sim}(s_i, s_i^{(-)})/\mu\right]}_{Y_{x,\eta} = f_Y(x, f_\eta(x))} - \mathbb{E}_{s_i \sim \mathcal{S}} \underbrace{\left[\text{sim}(s_i, s_i^{(+)})/\mu\right]}_{Y_{x,\eta^*} = f_Y(x, \eta_\emptyset)} \\ =& \mathbb{E}_{s_i \sim \mathcal{S}} \underbrace{[Y_{x,\eta} - Y_{x,\eta^*}]}_{TIE}. \end{aligned} \tag{10}$$

Analysis. The loss $\mathcal{L}_{ccl}$ can be transformed into the total TIE of $X = x$ on Y, which denotes the change of Y when only the $\mathcal{N}$ changes from η to $\eta*$ via a causal intervention $do(\mathcal{N} = \eta_\emptyset)$. We observed the underlying relations that (1) the expected noise-free state $\mathcal{N} = \eta_\emptyset$ can be utilized as a positive sample for CL, and (2) the way to reinforce cohesion in positive pairs and diversity in negative pairs in CL can serve as the prediction of Y in causal inference. In this way, our CCL takes full advantage of both causal inference and CL. The former excels at reducing the noise categories in the support samples for generating prototypes, preventing our model from overly predicting more undesired categories. The latter can facilitate more discrete and distinguishable category representations, thereby mitigating the semantic ambiguity issue. We then obtain denoised anchor representation s_i by optimizing the $\mathcal{L}_{ccl}$.

Dataset	#cls.	#inst./cls.	#inst.
FewAsp (Random)	100	630	63,000
FewAsp (Multi)	100	400	40,000

Table 1: Statistics of datasets. "#cls.", "#inst./cls.", and "#inst." indicate the number of classes, instances per class, and instances, respectively.

Different from recent works that regard the representations of other categories as negative pairs for contrast (Zhao et al., 2022; Liu et al., 2022), our CDPN involves extracting noise representations from anchor representations as negative pairs. We also found that the discrete noise representation works better by incorporating multiple negative labels, while the continuous noise representation can be obtained without requiring negative labels.

4.3 Support-Query Alignment

Following recent work (Peng et al., 2024), we leverage cross-attention mechanisms to emphasize shared semantic representations s^v across k instances of the target aspect category as follows:

$$\begin{aligned} s^v &= \frac{1}{K} \sum_{k=1}^{K} (\gamma_t^v)^{\top} h_{i,t}^k, \\ \gamma_t^v &= \text{softmax}(\cos(s_i^{\top} h_{i,t'}^k)). \end{aligned} \tag{11}$$

We apply cross-attention to the query embeddings and all support samples (e.g., $n \times k$) to obtain the query representation q^v. We utilize a simple shared mapping function (i.e., a linear transformation $\mathbf{FN}(\cdot)$) to align support and query representations, $\hat{s}_v$ and $\hat{q}_v$, which is given as follows:

$$\hat{s}_v = \mathbf{FN}([s^v]), \hat{q}_v = \mathbf{FN}([q^v]). \tag{12}$$

As such, we obtain the enhanced prototype p^n by substituting the s_i^n with $\hat{s}_v$ in Eq. (2). Similarly, we obtain the enhanced $\hat{q_v}$ to substitute q in Eq. (3).

4.4 Model Optimization

The model is optimized through two objectives, i.e., the MSE loss for the aspect category detection task and the CCL loss for anchor representation denoising, which is given by:

$$\mathcal{L}_{total} = \delta_1 \mathcal{L}_{acd} + \delta_2 \mathcal{L}_{ccl}, \tag{13}$$

where δ_1, δ_2 are hyperparameters.

Model	5-way 5-shot		5-way 10-shot		10-way 5-shot		10-way 10-shot	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Non-label-driven Approach								
PN (Snell et al., 2017)	88.88	66.96	91.77	73.27	87.35	52.06	90.13	59.03
IMP (Allen et al., 2019)	89.95	68.96	92.30	74.13	88.50	54.14	90.81	59.84
Proto-HATT (Gao et al., 2019)	91.54	70.26	93.43	75.24	90.63	57.26	92.86	61.51
Proto-AWATT (Hu et al., 2021)	93.35	75.37	95.28	80.16	92.06	65.65	93.42	69.70
Label-driven Approach								
LDF (Zhao et al., 2022)	94.65	78.27	95.71	81.87	92.74	67.13	94.24	71.97
LNP (Liu et al., 2022)	96.45	82.22	97.15	84.90	95.36	71.42	96.55	76.51
FSO (Zhao et al., 2023)	96.92	83.44	97.38	85.08	95.65	73.78	96.28	76.58
LGP (Guan et al., 2024)	97.37	87.49	97.49	87.67	96.33	77.92	96.69	78.95
VHAF (Peng et al., 2024)	97.88	87.25	98.17	89.22	97.02	79.72	97.58	82.41
CDPN	**98.09**	**88.26**	**98.55**	**90.42**	**97.76**	**83.07**	**98.12**	**84.58**

Table 2: Performance comparison on FewAsp (Random). **Bold**: Best, underline: Second best.

Model	5-way 5-shot		5-way 10-shot		10-way 5-shot		10-way 10-shot	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Non-label-driven Approach								
PN (Snell et al., 2017)	89.67	67.88	91.60	72.32	88.01	52.72	90.68	58.92
IMP (Allen et al., 2019)	90.12	68.86	92.29	73.51	88.71	53.96	91.10	59.86
Proto-HATT (Gao et al., 2019)	91.10	69.15	93.03	73.91	90.44	55.34	92.38	60.21
Proto-AWATT (Hu et al., 2021)	91.45	71.72	93.89	77.19	89.80	58.89	92.34	66.76
Label-driven Approach								
LDF (Zhao et al., 2022)	95.66	79.48	96.55	82.81	94.51	67.28	95.66	71.87
LNP (Liu et al., 2022)	92.62	73.38	94.34	78.81	90.87	62.06	92.93	68.23
FSO (Zhao et al., 2023)	96.01	81.04	96.67	82.22	94.93	70.26	95.71	72.46
LGP (Guan et al., 2024)	97.67	85.22	97.86	86.08	95.89	75.01	96.35	76.97
VHAF (Peng et al., 2024)	97.09	84.64	97.57	87.31	96.01	75.92	96.78	79.43
CDPN	**97.73**	**87.77**	**97.91**	**88.62**	**96.69**	**80.75**	**97.69**	**82.26**

Table 3: Performance comparison on FewAsp (Multi).

5 Experiments

5.1 Experimental Setting

Datasets and Metrics. We conducted experiments on two public datasets: FewAsp (Random) and FewAsp (Multi). The FewAsp (Multi) dataset consists of multi-aspect sentences, whereas the FewAsp (Random) dataset contains single- and multi-aspect sentences by random sampling. Both datasets contain 100 aspects, with 64 aspects for training, 16 for validation, and 20 for testing. The statistics of the datasets are shown in Table 1. We followed the same setting as Zhao et al. (2022)'s work to process data. We used Area Under the Curve (AUC) and Macro-F1 score (F1) as evaluation metrics.

Implementation. Following previous works (Zhao et al., 2022, 2023), we performed experiments with N={5, 10, 15} and K={2, 3, 5, 10}. The number of query instances per category was 5. The parameters are as follows: δ_1 and δ_2 were set to 1 and 0.1, respectively. The temperature μ was set to 0.1, 0.05, and 0.05 when N=5, 10, and 15, respectively. The initial learning rate was 2e-5. The weight decay was set to 1e-3. We randomly sampled 500 meta-tasks for training, the number of meta-tasks during the validation and testing was both set to 600, and the fixed threshold to select the positive category predictions in the 5-way setting, 10-way setting, and 15-way setting are set to 0.3, 0.2, 0.2, respectively. We reported the average testing results for 5 runs, where the seeds were set to [5, 10, 15, 20, 25]. The CDPN was implemented and experimented with PyTorch on Nvidia GeForce RTX 4090 (24GB memory). We utilized the AdamW optimizer (Loshchilov and Hutter, 2017).

Baselines. We compared CDPN with nine SOTA baselines in two groups.

Non-label-driven methods: PN (Snell et al., 2017), IMP (Allen et al., 2019), proto-HATT (Gao et al., 2019), and proto-AWATT (Hu et al., 2021).

Label-driven methods: LDF (Zhao et al., 2022), LPN (Liu et al., 2022), FSO (Zhao et al., 2023), LGP (Guan et al., 2024), and VHAF (Peng et al., 2024). Further details are provided in the Appendix A.1.

5.2 Experimental Results

Tables 2 and 3 show the performance comparison. The CDPN consistently outperforms all baselines on both datasets, which indicates the effectiveness of CDPN for few-shot MACD tasks. Specifically, CDPN surpasses the state-of-the-art baselines, LGP

Model	10-way 2-shot		10–way 3-shot		15-way 2-shot		15–way 3-shot	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
LDF (Zhao et al., 2022)	89.69	57.45	90.16	59.67	88.30	40.98	89.25	40.07
LNP (Liu et al., 2022)	91.36	56.11	92.67	60.34	90.88	48.24	92.21	53.38
FSO (Zhao et al., 2023)	93.97	66.41	94.32	68.28	93.30	59.25	93.69	62.85
VHAF (Peng et al., 2024)	94.99	70.69	96.08	75.56	95.20	67.89	95.12	70.22
CDPN w/o CCL	95.07	70.77	95.83	75.54	94.43	66.03	95.27	69.89
Full CDPN	**95.58**	**73.48**	**96.12**	**77.40**	**95.66**	**69.73**	**95.58**	**73.48**

Table 4: Performance comparison in scenarios with more categories and fewer shots on FewAsp (Multi).

Category		5-W 5-S		5-W 10-S		10-W 5-S		10-W 10-S	
CCL	SQA	AUC	F1	AUC	F1	AUC	F1	AUC	F1
✘	✘	95.9	81.7	96.1	82.3	94.6	69.4	95.9	74.3
✘	✔	96.1	85.6	97.3	85.3	95.5	76.3	96.2	78.5
✔	✘	97.3	85.8	97.6	86.7	96.5	76.1	96.6	77.6
✔	✔	**97.7**	**87.8**	**97.9**	**88.6**	**96.7**	**80.8**	**97.7**	**82.3**
Imp. (%)		**1.9**	**7.5**	**1.9**	**7.7**	**2.3**	**16.3**	**1.9**	**10.7**

Table 5: Ablation study on FewAsp (multi). "SQA" indicates the support-query alignment. ✔ and ✘ denote with and without each module, respectively.

and VHAF, with the improvement of 0.21% $\sim$ 0.76% AUC, and 0.88% $\sim$ 4.20% F1 on FewAsp (Random). On FewAsp (multi), CDPN leads a performance boost of 0.05% $\sim$ 0.94% and 1.50% $\sim$ 6.36% in AUC and F1, respectively. We also have the following observations:

(1) LDF and LNP employ contrastive learning by using other categories as negative pairs. Conversely, CDPN derives noise representations from anchor representations as negative pairs, achieving better performance. This highlights the superiority of our CCL in this task.

(2) VHAF, which utilizes variational distribution inference to address the distribution shift between the support set and query set, stands out as a strong performer among the baselines. In contrast, our CDPN, by denoising support samples and support-query alignment, performs better.

(3) Overall, FewAsp (multi) is more challenging than FewAsp (Random) for MACD tasks, as evidenced by the lower AUC and F1 values across all methods. One plausible reason is that samples in the FewAsp (multi) dataset are exclusively multi-aspect sentences, whereas sentences in the FewAsp (Random) dataset can be either single-aspect or multi-aspect, resulting in fewer noise categories.

5.3 Ablation Study

We conducted an ablation study to examine the effects of each component in CDPN. As shown in Table 5, we have the following observations:

(1) The CCL and SQA modules improve performance by 1.9% to 2.3% in AUC and 7.5% to 10.7% in F1, respectively. The improvement is more significant with a higher number of ways, e.g., the F1 improvement in 10-way 5-shot compared to 5-way 5-shot. Overall, CCL contributes more to the performance of CDPN than SQA.

(2) Without the CCL module (i.e., w/o CCL), CDPN suffers from a significant performance drop in F1 for both datasets, demonstrating that integrating causal inference and contrastive learning can effectively denoise support samples for boosting the prototype representations.

(3) The underperformance of CDPN without representation alignment (i.e., without SQA) suggests that it is important to emphasize shared semantic representations across instances for the target aspect category.

5.4 Performance of CDPN with More Categories and Fewer Shots

We compared CDPN with other baselines with more categories and fewer support samples (N=10 and 15 and K=2 and 3). As shown in Table 4, most baselines suffer from a significant performance drop as N increases to 10 and 15, whereas CDPN achieves an improvement of 2.44% to 4.64% in F1 compared to the best baseline. The VHAF model demonstrates impressive performance compared to other baselines because they statistically recalibrate the distribution of support and query samples. In contrast, CDPN exhibits greater robustness to noise categories with fewer support samples. This is due to our CCL module, i.e., the result (w/o CCL) in Table 4 shows a significant performance decrease.

5.5 Performance of Various Noise

We conducted experiments to explore how discrete and continuous noises affect the CDPN. As shown in Fig 4, we have two observations:

(1) The discrete noise representation works better than the continuous noise representation, which demonstrates that the discrete noise representation facilitates better convergence for CDPN by incor-

Model	5-W 2-S		10-W 2-S		5-W 2-S		10-W 2-S	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
	FewAsp (random)				FewAsp (multi)			
Llama-2-7B	73.74	47.06	68.57	35.42	71.25	46.67	70.76	35.31
Llama-3-8B	86.19	71.32	84.62	61.12	84.47	67.87	83.75	59.75
GPT-3.5	89.30	82.11	89.15	70.68	88.83	77.53	86.78	67.58
GPT-4	91.47	82.12	89.15	74.37	88.63	79.97	86.56	71.86
CDPN (0.2B)	**97.39** (↑6.5%)	**85.89** (↑4.6%)	**96.74** (↑8.6%)	**76.39** (↑2.7%)	**96.31** (↑8.8%)	**83.65** (↑4.6%)	**95.58** (↑10.1%)	**73.48** (↑2.2%)

Table 6: Performance comparison with Llama and GPT models. Due to their unacceptable time consumption compared with our CPDN, we only report their results with N=5 and 10, and K=2.

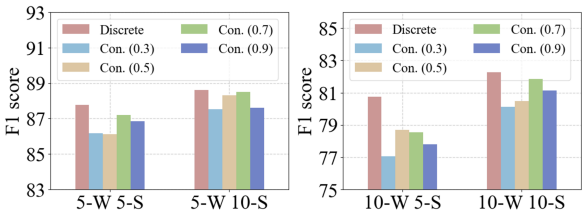


Figure 4: Performance of various noise representations on FewAsp (Multi). "Discrete" indicates discrete noise. " Con. (r)" refers to continuous noise at dropout rate r.

porating multiple negative labels.

(2) For continuous noise representation, the higher dropout rate (i.e., 0.7), produces better results than lower ones (i.e., 0.3 and 0.5), because the representations with a lower dropout rate are not distinguishable from the anchor representations. Notably, the advantage of continuous noise representation lies in its ability to be obtained without the requirement for negative labels.

5.6 Comparison with LLMs in MACD Tasks

We compared CDPN with Llama-2-7B, Llama-3-8B, GPT-3.5-turbo, and GPT-4o-mini in Table 6. We have the following findings:

(1) The results in Table 6 show that both Llama-2-7B and Llama-3-8B still have challenges with domain-specific MACD tasks. This is because each support sample often contains multiple noisy categories, confusing them to generate irrelevant inferences instead of predicting categories. In contrast, as a small task-specific model, CDPN (i.e., 0.2B) achieves better performance in MACD tasks.

(2) Although the Llama models (7B and 8B) perform well in most cases due to the extensive knowledge, they still have two challenging cases.

Case (1):
Did I mention they even had a vegan dessert?

Ground truth: "food_dessert"
Llama: "food_dessert" ✓ "food_mealtype_lunch" ×
CDPN: "food_dessert" ✓

The Llama model tends to overly infer more aspect categories, while our model, by denoising prototypes, can mitigate spurious correlations. For example, Llama models incorrectly predict "food_mealtype_lunch" along with the correct category "food_dessert." In contrast, our model effectively mitigates these spurious correlations by employing a prototype denoising approach.

Case (2):
We didn't have to fight for chairs like other cheap places I've been to such as hard rock or MGM grand.

Ground truth: "room_interior"
Llama: "room_overall" × "food_food"×
CDPN:" "room_interior" ✓

Some implicit aspect categories confuse Llama models, causing them to produce hallucinations. For example, Llama-3-8B incorrectly includes "food_food" as a predicted category. When we asked why to include "food_food", the Llama model explained that "it can be inferred the speaker is comparing the food quality or service at this location to those at Hard Rock and MGM Grand."

(3) We compared CDPN with GPT-3.5-turbo and GPT-4o-mini. In task-specific MACD, CDPN achieves competitive results by denoising category prototypes. The GPT-3.5-turbo and GPT-4o-mini excel in discovering subtle associations among words (e.g., once the name of a person appears, they assume that "staff" is mentioned), leading them to predict more categories. Therefore, multi-category samples are easier to predict, while CDPN and smaller LLMs better identify random-category samples (a mix of single and multiple categories) than multi-category samples.

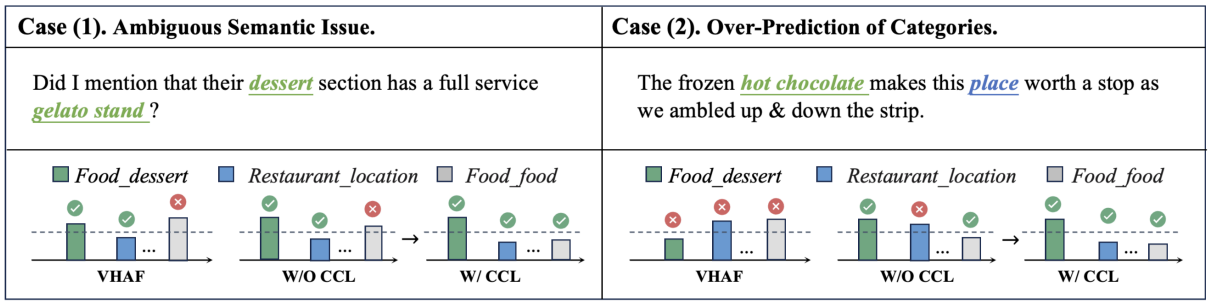


Figure 5: Case study. Dashed lines indicate the predicted AC threshold, consistent with the setup in (Zhao et al., 2022). In case (1), "Food_food" is an irrelevant category but wrongly detected. In case (2), "place" leads models to mistakenly select "Restaurant_location", while the reviewer intends to refer to the dessert, not the location.

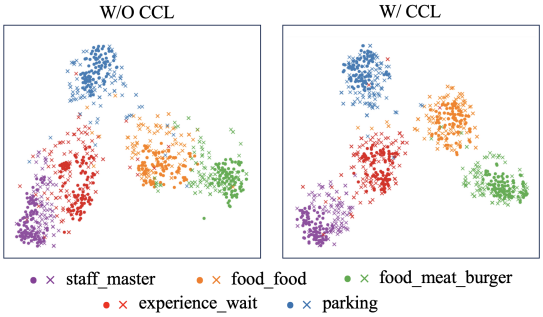


Figure 6: Visualization using t-SNE (van der Maaten and Hinton, 2008). Dots (•) and crosses (×) represent prototype and query representations, respectively.

5.7 Case Study and Visualization

Fig. 5 presents two cases of CDPN with and without CCL, as well as the baseline VHAF:

Case (1). The ambiguous semantic issue is often challenging in MACD tasks. It can be observed that the use of CCL enables CDPN to better distinguish similar categories, such as "food_food" and "food_dessert," whereas both CDPN without CCL and VHAF failed to make correct predictions.

Case (2). The model without CCL tends to overly infer more categories. This may be because "place" often appears in the same context as "Restraurant_location", leading to their over-association. The CCL can mitigate this issue.

Visualization. We visualized category prototype and query representations with and without CCL in Fig. 6. We have two findings: (1) the CCL enables CDPN to learn more distinguishable representations among categories (i.e., "staff_master" and "experimence_wait", and "food_food" and "food_meat_burger"), and (2) with CCL, the distributions of prototype and query representations exhibit a higher degree of overlap.

6 Conclusion

We proposed a concise task-specific CDPN model for the few-shot MACD tasks. We integrated causal inference and contrastive learning to eliminate the undesired impact of noise categories in the support set. We explored discrete and continuous noise to build negative pairs for CCL. Extensive experiments have demonstrated that our CDPN achieves SOTA performance. In future work, we intend to (i) generate more counterfacts with LLMs for enhancing CDPN and (ii) examine our CCL aspect sentiment triplet extraction (ASTE) (Gou et al., 2023) and opinion aspect target sentiment quadruple extraction (OATS) (Chebolu et al., 2024).

Limitations

The proposed causal contrastive learning can only detect the aspect categories, while aspect terms, opinions, and sentiment polarities are all important for aspect-based sentiment analysis tasks. We will extend our approach to a wider variety of ABSA tasks (Gou et al., 2023; Cui et al., 2023, 2024). Another limitation is that the CCL is primarily designed for denoising tasks, which limits its generalization to more common applications.

Ethics Statement

This paper does not involve the presentation of a new dataset, an NLP application, or the use of demographic or identity characteristics information.

Acknowledgements

We would like to thank anonymous reviewers for their thorough comments and suggestions. This work is supported by JKA (2024M-557), JSPS KAKENHI (No. 24K15085), SCAT, and the China Scholarship Council (No.202208330091).

References

Kelsey Allen, Evan Shelhamer, Hanul Shin, and Joshua Tenenbaum. 2019. Infinite mixture prototypes for few-shot learning. In *International conference on machine learning*, pages 232–241. PMLR.

Boxi Cao, Hongyu Lin, Xianpei Han, Fangchao Liu, and Le Sun. 2022. Can prompt probe pretrained language models? understanding the invisible risks from a causal view. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5796–5808.

Mingshan Chang, Min Yang, Qingshan Jiang, and Ruifeng Xu. 2024. Counterfactual-enhanced information bottleneck for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17736–17744.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2024. OATS: A challenge dataset for opinion aspect target sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12336–12347, Torino, Italia. ELRA and ICCL.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2021. Honey or poison? solving the trigger curse in few-shot event detection via causal intervention. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088.

Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. 2022. Contrastnet: A contrastive learning framework for few-shot text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10492–10500.

Xiudi Chen, Hui Wu, and Xiaodong Shi. 2023a. Consistent prototype learning for few-shot continual relation extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7409–7422.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023b. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638.

Jin Cui, Fumiyo Fukumoto, Xinfeng Wang, Yoshimi Suzuki, Jiyi Li, and Wanzeng Kong. 2023. Aspect-category enhanced learning with a neural coherence model for implicit sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11345–11358.

Jin Cui, Fumiyo Fukumoto, Xinfeng Wang, Yoshimi Suzuki, Jiyi Li, Noriko Tomuro, and Wanzeng Kong. 2024. Enhanced coherence-aware network with hierarchical disentanglement for aspect-category sentiment analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5843–5855.

Fuli Feng, Weiran Huang, Xiangnan He, Xin Xin, Qifan Wang, and Tat-Seng Chua. 2021. Should graph convolution trust neighbors? a simple causal inference method. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1208–1218.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6407–6414.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. Mvp: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397.

ChaoFeng Guan, YaoHui Zhu, Yu Bai, and LingYun Wang. 2024. Label-guided prompt for multi-label few-shot aspect category detection. *arXiv preprint arXiv:2407.20673*.

Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A generative language model for few-shot aspect-based sentiment analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 770–787.

Mengting Hu, Shiwan Zhao, Honglei Guo, Chao Xue, Hang Gao, Tiegang Gao, Renhong Cheng, and Zhong Su. 2021. Multi-label few-shot learning for aspect category detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6330–6340.

Sabyasachi Kamila, Walid Magdy, Sourav Dutta, and MingXue Wang. 2022. Ax-mabsa: A framework for extremely weakly supervised multi-label aspect based sentiment analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6136–6147.

Kexin Li, Chengjiang Long, Shengyu Zhang, Xudong Tang, Zhichao Zhai, Kun Kuang, and Jun Xiao. 2024. Corerec: A counterfactual correlation inference for next set recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8661–8669.

Yuncong Li, Cunxiang Yin, Sheng-hua Zhong, and Xu Pan. 2020. Multi-instance multi-label learning networks for aspect-category sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3550–3560.

Bin Liang, Xiang Li, Lin Gui, Yonghao Fu, Yulan He, Min Yang, and Ruifeng Xu. 2023. Few-shot aspect category sentiment analysis via meta-learning. *ACM Transactions on Information Systems*, 41(1):1–31.

Bin Liang, Qinlin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. Jointcl: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 81–91. Association for Computational Linguistics.

Han Liu, Feng Zhang, Xiaotong Zhang, Siyang Zhao, Junjie Sun, Hong Yu, and Xianchao Zhang. 2022. Label-enhanced prototypical network with contrastive learning for multi-label few-shot aspect category detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1079–1087.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. 2020. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*.

Venkata Prabhakara Sarath Nookala, Gaurav Verma, Subhabrata Mukherjee, and Srijan Kumar. 2023. Adversarial robustness of prompt-based few-shot learning for natural language understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2196–2208, Toronto, Canada. Association for Computational Linguistics.

Cheng Peng, Ke Chen, Lidan Shou, and Gang Chen. 2024. Variational hybrid-attention framework for multi-label few-shot aspect category detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 14590–14598.

Chen Qian, Fuli Feng, Lijie Wen, Chunping Ma, and Pengjun Xie. 2021. Counterfactual inference for text classification debiasing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5434–5445.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. 2023. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*.

Teng Sun, Wenjie Wang, Liqaing Jing, Yiran Cui, Xuemeng Song, and Liqiang Nie. 2022. Counterfactual reasoning for out-of-distribution multimodal sentiment analysis. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 15–23.

Bing Tian, Yixin Cao, Yong Zhang, and Chunxiao Xing. 2022. Debiasing nlu models via causal intervention and counterfactual reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11376–11384.

Geng Tu, Ran Jing, Bin Liang, Min Yang, Kam-Fai Wong, and Ruifeng Xu. 2023. A training-free debiasing framework with counterfactual reasoning for conversational emotion detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15639–15650.

Can Udomcharoenchaikit, Wuttikorn Ponwitayarat, Patomporn Payoungkhamdee, Kanruethai Masuk, Weerayut Buaphet, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. Mitigating spurious correlation in natural language understanding with counterfactual inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11308–11321.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. *Advances in neural information processing systems*, 34:16196–16208.

Siyin Wang, Jie Zhou, Changzhi Sun, Junjie Ye, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2022. Causal intervention improves implicit sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6966–6977.

Wenjie Wang, Fuli Feng, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1288–1297.

Xinghao Wang, Junliang He, Pengyu Wang, Yunhua Zhou, Tianxiang Sun, and Xipeng Qiu. 2024. Denosent: A denoising objective for self-supervised sentence representation learning. *arXiv preprint arXiv:2401.13621*.

Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Zengzhi Wang, Qiming Xie, and Rui Xia. 2023. A simple yet effective framework for few-shot aspect-based sentiment analysis. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1765–1770.

Zeyu Wang and Mizuho Iwaihara. 2023. Few-shot multi-label aspect category detection utilizing prototypical network with sentence-level weighting and label augmentation. In *International Conference on Database and Expert Systems Applications*, pages 363–377. Springer.

Zhao Wang and Aron Culotta. 2021. Robustness to spurious correlations in text classification via automatically generated counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14024–14031.

Jialong Wu, Linhai Zhang, Deyu Zhou, and Guoqiang Xu. 2024. Diner: Debiasing aspect-based sentiment analysis with multi-variable causal inference. *arXiv preprint arXiv:2403.01166*.

Weizhi Xu, Qiang Liu, Shu Wu, and Liang Wang. 2023. Counterfactual debiasing for fact verification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6777–6789.

Jianhua Yuan, Yanyan Zhao, and Bing Qin. 2022. Debiasing stance detection models with counterfactual reasoning and adversarial bias learning. *arXiv e-prints*, pages arXiv–2212.

Haoxing Zhang, Xiaofeng Zhang, Haibo Huang, and Lei Yu. 2022a. Prompt-based meta-learning for few-shot text classification. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 1342–1357.

Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 2021. Causerec: Counterfactual user sequence synthesis for sequential recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 367–377.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022b. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

Fei Zhao, Yuchen Shen, Zhen Wu, and Xinyu Dai. 2022. Label-driven denoising framework for multi-label few-shot aspect category detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2390–2402.

Shiman Zhao, Wei Chen, and Tengjiao Wang. 2023. Learning few-shot sample-set operations for noisy multi-label aspect category detection. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5306–5313.

Yangqiaoyu Zhou, Yiming Zhang, and Chenhao Tan. 2023. FLamE: Few-shot learning from natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6743–6763, Toronto, Canada. Association for Computational Linguistics.

A Appendix

This section provides further implementation setups and experimental results.

A.1 Baselines

We validate the effectiveness of the CDPN model by comparing it with the following eight baselines:

- ***Non label-driven methods:***

- Prototypical Network **(PN)** (Snell et al., 2017) averages the support sample representations as the prototype, and measures the distance between query instances and each prototype.
- Infinite Mixture Prototypes **(IMP)** (Allen et al., 2019) attempts to infinite mixture prototypes to represent each category for few-shot learning.
- Proto-HATT **(HATT)** (Gao et al., 2019) uses a hybrid instance-level and feature-level attention mechanism to enhance instance features.
- Proto-AWATT **(AWATT)** (Hu et al., 2021) utilizes support-set attention and query-set attention to alleviate the noise.

- ***Label-driven methods:***

- **LDF** (Zhao et al., 2022) uses a label-weighted contrastive loss as the complementary objective function and label-guided attention to filter out noisy words, and generates a representative prototype for each category.
- **LPN** (Liu et al., 2022) utilizes the descriptions of label words as the label information, and employs multi-head self-attention and contrast learning to enhance prototypes.
- **FSO** (Zhao et al., 2023) employs a group of sample-set operations to perform prototype denoising for MACD.
- **LGP** (Guan et al., 2024) proposes a label-guided prompt method to obtain sentence representations and category prototypes.
- **VHAF** (Peng et al., 2024) introduces a variational distribution inference model to estimate the output of query samples based on the distribution of support samples.

A.2 Robustness Study

To further show the robustness of CDPN, we obtained the variance (Var.) and a 10-trial T-test on the FewAsp (Random) and FewAsp (Multi) datasets and reported the results in Table 7. "*" indicates that the improvement of our CDPN is statistically significant compared with the second-best. In all cases, the improvement compared with the second-best baselines is statistically significant (p-value < 0.05).

Metric	5-W 5-S	5-W 10-S	10-W 5-S	10-W 10-S
FewAsp (Random)				
SOTAs	87.49	89.22	79.72	82.41
Avg.	88.26*	90.42*	83.07*	84.58*
Var.	0.69	0.41	0.18	0.13
p-value:	4.1e-4	2.5e-7	1.1e-13	4.9e-11
FewAsp (Multi)				
SOTAs	85.22	87.31	75.92	79.43
Avg.	87.77*	88.62*	80.75*	82.26*
Var.	0.23	0.33	2.1	0.17
p-value:	4.9e-11	2.0e-5	8.0e-07	7.2e-14

Table 7: Robustness study in terms of F1 on the FewAsp (Random) and FewAsp (Multi) datasets. "SOTAs" refers to the second-best baselines, i.e., LGP and VHAF.

A.3 Visualization

We visualized distributions of prototype and query representations learned by CDPN in Fig. 7. We randomly selected 5, 10 and 15 aspects from the test set of the Fewasp (multi) dataset. We have the following findings and insights:

(1) Overall, the same colors are clustered together, indicating that the CDPN effectively identifies similar prototypes and query representations. Additionally, it is evident that more ways and fewer shots make the MACD more challenging.

(2) We observed that some dots or crosses in different colors are close to each other. There are two reasons for this: (i) the majority of sentences in the support and query sets involve multiple aspect categories; and (ii) many categories are semantically similar, e.g., "food_dessert,""food_burger," and "food_mealtype_lunch."

A.4 Prompt Templates for Llama models

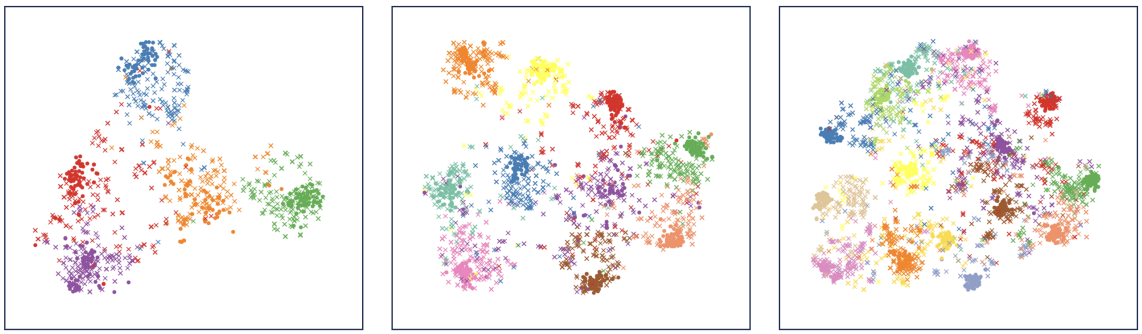
We compared CDPN with Llama-2-7B and Llama-3-8B on two datasets. The 5-way 2-shot results generated by Llama models were produced using the following template:

Prompt Template for Llama:
The few-shot multi-label aspect category detection task refers to detecting aspect categories within a given sentence assigned multiple categories with minimal annotated examples.

Here are two examples for each aspect category as support sentences: {*support examples*}

The available multiple categories include the following 5 selections: {*aspect selections*}

Given a test sentence, please perform the few-shot multi-label aspect category detection task, and directly return the label without other text.



5-way 2-shot 10-way 2-shot 15-way 2-shot

Figure 7: Visualization of various prototypes. Dots (•) and crosses (×) represent prototype and query representations, respectively. Different colors represent different aspect categories.

{*support examples*}:

"parking":
(1) He was with our dog in the car but thanks.
(2) The MGM Grand parking is plentiful and pretty easy to navigate.

"food_mealtype_lunch":
(1) They continued to make sure everything was good throughout our lunch.
(2) Pros: cheap lunch buffet, convenient location for CMU students, good for me cons: not great for those looking for superspicey dishes star of India reminds me of that one summer in Pittsburgh the one when we worked all the time and only had a weekly break ... to come down to Craig street for the lunch buffet.

"room_bed":
(1) With a full kitchen, washer and dryer, patio, or balcony and TV in every bedroom.
(2) Comfortable beds, cheap prices but still feels luxurious.

"room_smoke":
(1) Not unless you want a smoking room.
(2) The overpowering smell of smoke and lord knows what else was enough to make me queasy.

"experience_wait":
(1) In the 30 or more visits I've had, my average wait time has been maybe 2 minutes.
(2) I can't wait to go back.

{*aspect selections*}:

"**parking**", "**food_mealtype_lunch**", "**room_bed**", "**room_smoke**", "**experience_wait**"