Explicit Modeling of Causal Factors and Confounders for Image Classification

Anonymous submission

Abstract

Causal inference has emerged as a promising approach for identifying decisive semantic factors and eliminating spurious correlations in visual representation learning. However, most existing methods rely on latent, data-driven confounder modeling, normally attributing the source of bias to background information while neglecting object-level semantic confusions that commonly occur in complex scenes. This limits their effectiveness in disentangling causal factors from confounding semantics. To address this challenge, we propose an explicit modeling approach for both causal factors and confounders, termed Explicit Modeling Causal Model (EMCM). The proposed framework consists of three key components. The Features Stability Estimation module explicitly models the relationship between visual semantics and class labels by leveraging clustering patterns to perform class-aware separation of causal and confounding factors. It produces class-specific causal factors and confounding factors linked to ambiguous categories. Subsequently, the Discriminative Features Enhancing module integrates causal factors into fused patch features via front-door intervention for stable semantics. In parallel, the Explicit Confounder Modeling and Debiasing Module learns confounders under clear label guidance and derives debiased context features by TDE modeling. This framework leverages two complementary causal perspectives to construct a unified semantic representation that facilitates improved generalization. Extensive experiments on two datasets demonstrate that EMCM effectively disentangles causal and confounding factors in complex scenarios, consistently outperforming state-of-the-art causal debiasing methods and text-guided methods in all metrics.

Introduction

Text-enhanced image classification aims to utilize more discriminative textual information to guide visual representation learning, solving challenges posed by inter-class similarity and intra-class diversity in visual data. Existing methods typically enforce the alignment of heterogeneous modality representations to adjust the distribution of high-dimensional visual features and enhance identifiability. However, as the complexity of visual data increases, the performance of cross-modal learning deteriorates. This is due to redundant features and spurious correlations that are common in images, restricting the effectiveness of alignment.

Naive methods can be categorized into two types: feature-level alignment methods and fine-grained alignment methods. The former typically maps features from different modalities directly into a shared latent space (Guan et al. 2023; Meng et al. 2019) and employs regularizations (Radford et al. 2021; Yang et al. 2022) to bring the representations of corresponding texts and images closer while separating the unrelated pairs. However, heavy compression of semantic information in global features degrades alignment performance. To address this, the latter typically leverages detection models to extract object regions, aligning specific visual elements with words to exclude the interference of irrelevant elements (Wei et al. 2020; Pan, Wu, and Zhang 2023). However, these purely data-driven methods can only capture statistical correlations and fail to distinguish between visual elements that are correlated but not causally related. To further mitigate the influence of spurious correlations, causal inference methods have been introduced to model the associations among visual components, enabling the identification of causal factors and confounders. These methods aim to suppress features detrimental to downstream tasks through interventions and counterfactual reasoning.

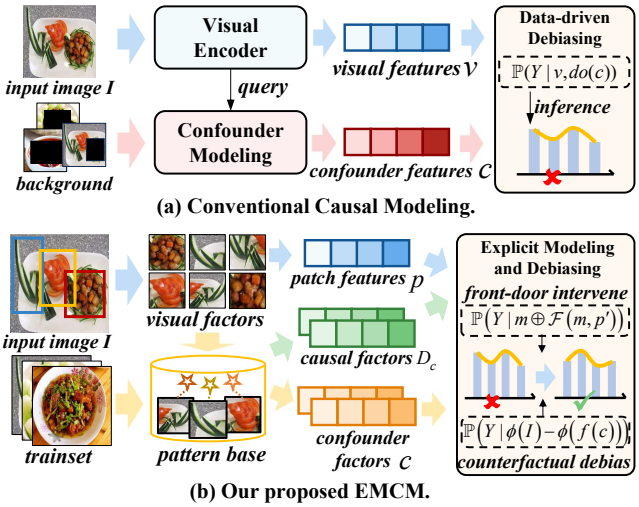

Figure 1: In contrast to conventional causal methods that model confounders in a data-driven manner, our proposed EMCM leverages a visual pattern base to facilitate the explicit modeling of causal and confounding features.

Despite the integration of causal inference into representation learning, existing frameworks still suffer from several issues: (1) Confounder extraction predominantly focuses on coarse foreground-background distinctions, which leads to insufficient disentanglement granularity in complex multi-object scenarios, e.g., Chinese food recognition; (2) Most of the existing methods construct latent confounder features in a data-driven manner from a predefined confounder pool, yet they lack explicit semantic-level guidance, which limits their generalization ability.

To address these challenges, this paper presents an Explicit Modeling Causal Model (EMCM) that extracts causal and confounding factors by analyzing the pre-defined semantic pattern base, while forming synergistic representations by fusing causal stable features with debiased contextual features for better generalization. Specifically, EMCM involves three key modules: Feature Stability Estimation (FSE) module, Discriminative Feature Enhancing (DFE) module, and Explicit Confounder Modeling and Debiasing (ECMD) module. Specifically, the FSE module utilizes a clustering-pattern base, consisting of semantically intensive yet noisy visual patches extracted by the pre-trained Grounding DINO model, to explicitly separate stable causal factors from unstable confounding factors. This arises from the observation that visual factors shared across multiple categories tend to be confounders, while those unique to a specific category are likely to become discriminative cues. By analyzing the purity of categories in clusters, the FSE outputs causal and confounding factors as well as confusing labels. The DFE module is applied to integrate causal factors with patch features through front-door intervention. In parallel, the ECMD module refines confounders and obtains context-debiased global features via counterfactual inference. EMCM achieves better generalization by mitigating the influence of contextually spurious correlations and emphasizing truly discriminative causal information.

Extensive experiments have been conducted on the VireoFood-172 and NUS-WIDE datasets to demonstrate the superiority of EMCM, including performance comparison with SOTA methods, ablation study, in-depth analyses, and case studies. The results validate the effectiveness of the EMCM in explicitly uncovering causal factors and confounders through semantic relationships and in combining stable causal features with debiased contextual information to enhance generalization. The contributions are as follows:

- This paper proposes a framework that explicitly models the causal and confounding factors for image classification. It is achieved by examining the relationship between category distributions and visual patterns for feature stability estimation. To the best of our knowledge, this is the first attempt to incorporate object-level semantics to facilitate explicit causal modeling.
- This paper adaptively constructs sample-specific confounder features based on confounding patterns and category constraints, ensuring consistent debiasing across similar samples while simplifying the learning process, offering a novel approach for confounder feature extraction.
- Extensive experiments on two public datasets demonstrate that our explicit causal modeling approach outperforms existing latent confounder extraction methods by more effectively identifying spurious semantics in complex visual scenarios. In addition, in-depth analysis shows that our method is adaptive to low-quality and noisy patches that may arise from the pre-trained DINO model.

Related Work

Visual-textual alignment can be categorized into global alignment and fine-grained alignment. The former directly maps sample features to the same latent space (Meng et al. 2019; Guan et al. 2023). To achieve this, (Meng et al. 2019) adopts partial heterogeneous transfer to make shared information interact between modalities. More advanced clip-based methods (Radford et al. 2021; Andonian, Chen, and Hamid 2022) use contrastive loss to obtain better consistent associations from large paired datasets and perform well in zero-shot conditions. Meanwhile, fine-grained alignment (Pan, Wu, and Zhang 2023; Xie et al. 2022; Gao et al. 2024) requires local elements, such as text words and image ROI regions, to be paired individually, aspiring for better interpretability and stability. (Pan, Wu, and Zhang 2023) discovers the shared semantics of image and text by mining the informative region-word pairs and rejecting irrelevant alignments. (Gao et al. 2024) introduces softened targets derived from fine-grained intra-modal self-similarity, effectively incorporating local similarities and modeling many-to-many relationships across modalities. In contrast to the aforementioned purely data-driven approaches that perform implicit alignment, our method explicitly models causal and confounding factors based on the visual semantic distribution and leverages two complementary causal mechanisms to obtain more generalized representations.

Meanwhile, causal inference (Pearl, Glymour, and Jewell 2016) has gained increasing traction for its ability to remove data bias in multimedia tasks, e.g., image classification (Yang et al. 2023; Liu et al. 2022b; Wu et al. 2024b), video question answering (Zang et al. 2023; Liu, Li, and Lin 2023; Zhang, Zhang, and Xu 2023), image-text retrieval (Liu et al. 2024; Zhang et al. 2024; Li et al. 2023), image captioning (Yang et al. 2021; Liu et al. 2022a; Chen et al. 2024), etc. (Yang et al. 2023) investigates the adverse context bias of the datasets and proposes a plug-in causal intervention module based on backdoor adjustment. (Wu et al. 2024b) reveals two biases behind the attention supervision and reduces them by subtracting direct causal effects from total causal effects. To handle the biases in multimodal tasks, (Liu, Li, and Lin 2023) introduces a linguistic backdoor causal intervention module and a local-global front-door causal intervention module to mitigate the linguistic and visual spurious correlations, respectively. (Liu et al. 2024) aims to learn causally-invariant visual representations for cross-modal retrieval, satisfying the independence and sufficiency properties. Although these methods have shown certain effectiveness, they normally rely on foreground-background separation for latent modeling of confounders, lacking the ability to capture object-level spurious correlations in complex scenes. In contrast, EMCM performs explicit causal modeling at the object level, enabling the identification of class-discriminative semantics from rich and cluttered contexts.

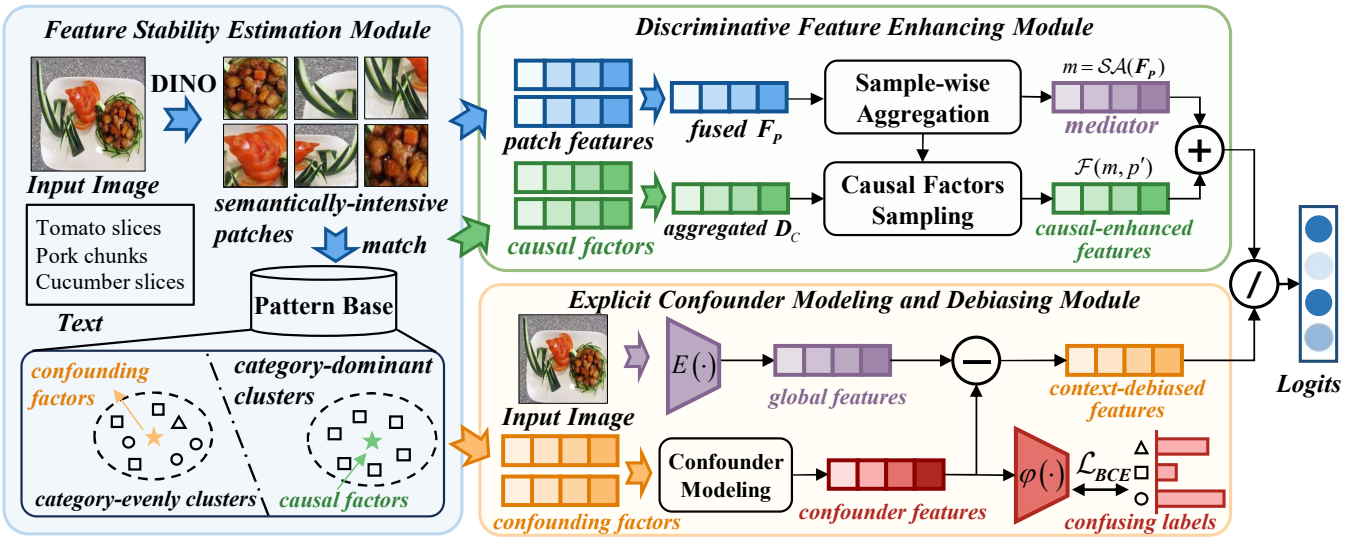

Figure 2: The overall architecture of EMCM. It separates causal from confounding factors via a pre-defined pattern base, then in parallel, enhances stable features using the causal factors while explicitly modeling and debiasing the confounder features.

Problem Formulation

We formulate the task under the Learning Using Privileged Information (LUPI) framework (Meng et al. 2019), where images $\{I_i\}_{i=1}^N$, labels $\{y_i\}_{i=1}^N$, and associated texts $t_{i\,j}^j$ are available during training, but only images are used at inference time. The objective is to learn consistent features for images sharing the same label. This is formulated as: $f_\theta(I_i) \rightarrow y_i$. To align visual features with textual semantic features, conventional causal methods model confounders and intervene in a data-driven manner. The process is represented as: $f_\theta(I_i, do(C)) \rightarrow y_i$, where C is the dataset-level latent confounders.

Different from traditional causal methods, EMCM first extracts semantically-intensive patches $\{p_i^j\}_{j=1}^{n_i}$ from image I_i via a pre-trained vision-language model. A visual encoder then captures patch-level features aggregated as F_P and a global image representation F_G. Then, the FSE module separates causal and confounding factors based on cluster category purity, yielding an aggregated causal dictionary D_c and $\mathrm{Confounder}(I_i)$. Subsequently, the DFE module facilitates causal reasoning by applying front-door intervention to obtain category-specific causal semantic features F_S. Meanwhile, the ECMD module refines confounder candidates with explicit constraints and then uses TDE modeling to extract debiased contextual features F_C. The process is formulated as: $f_\theta(F_{p_i}, D_c) \oplus f_\theta(I_i - \mathrm{Confounder}(I_i))) \rightarrow y_i$.

Methodology

The Structural Causal Model (SCM) (Pearl, Glymour, and Jewell 2016) is introduced to model the relationships among variables in image classification. As illustrated in Fig.3(a), the causal graph comprises five variables: image patches P, overall semantic content S, contextual information C, prediction Y, and confounder Z derived from the spurious semantic correlations and contextual noise. Two confounding paths are detected in this graph:

Path $P \leftarrow S \rightarrow Y$. When using DINO for visual element discretization, it essentially extracts semantically intensive visual regions. However, due to inherent dataset biases, object co-occurrence is likely to occur given specific semantic contexts. This results in a backdoor path between visual patches and the predicted label Y.

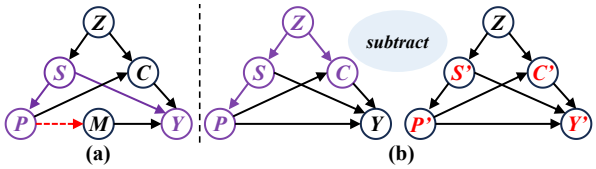

Figure 3: Causal view of the proposed EMCM.

Path $P \leftarrow S \leftarrow Z \rightarrow C$. Relying solely on object-level features while ignoring contextual cues empirically results in suboptimal classification performance (Liu et al. 2022b), particularly in complex scenarios where generalization is critical. Nevertheless, acquiring debiased contextual information is non-trivial, as backdoor pathway still exists.

This paper introduces EMCM, which explicitly models causal and confounding factors based on the DINO-produced semantically-intensive yet noisy visual patches. It consists of Feature Stability Estimation (FSE) module, Discriminative Feature Enhancing (DFE) module and Explicit Confounder Modeling and Debiasing (ECMD) module, see Figure 2. Specifically, FSE constructs a clustering-based visual pattern base and distinguishes patches into causal factors and confounder candidates. Subsequently, DFE employs front-door intervention to focus the model on causal features with stable discriminability, which essentially function as the mediator M. Meanwhile, ECMD explicitly refines confounder candidates under the constraint of confusing categories, and obtains debiased contextual features through counterfactual reasoning. The EMCM achieves a comprehensive causal understanding by integrating stable features with context-debiased representations.

Features Stability Estimation via Pattern Base

It is crucial to decouple the patch candidates by distinguishing between stably discriminative features and confounding features before conducting causal inference. Rather than

constructing a global confounder pool and extracting confounders in a data-driven manner (Liu et al. 2022b; Yang et al. 2023; Liu, Li, and Lin 2023), CCEM leverages a pattern base that models the purity of categories in each cluster to distinguish between causal and confounding visual patches. This promotes more precise identification of whether a specific visual pattern benefits the classification or functions as a confounder.

First, the advanced Grounding DINO (Caron et al. 2021) is used to extract semantically-intensive yet noisy patches from images, formulated as:

$$\{p_i\} = \mathrm{DINO}(I_i, \varphi_{text}(I_i)) \tag{1}$$

where φ_{text} is a pre-trained predictor for text prompt. The extraction process is performed offline, thereby not increasing the complexity of the inference process.

For patches from the training set, the K-means algorithm is applied to group patches into K clusters: $\mathbf{C}, \mathbf{Y} = \text{K-means}(\{P_i\}_{i=1}^{N})$, where $\mathbf{C} = \{C_i\}_{i=1}^{K}$ denotes cluster patterns, $\mathbf{Y} = \{y_j\}_{j=1}^{\sum_i n_i}$ denotes the cluster label, $P_i = \{p_i^j\}_{j=1}^{n_i}$ denotes the set of patches of image I_i. Subsequently, the purity of categories in each cluster is used to model causal factors and confounders simultaneously.

Specificallly, the distribution of categories within each cluster C_i is represented by the vector of proportions $\mathbf{y_i} = [p(y_1|C_i), ..., p(y_n|C_i)]$. Then, the purity of cluster C_i is defined as the maximum of proportions:

$$\mathrm{Purity}(C_i) = \max_{y \in Y} p(y|C_i) \tag{2}$$

To determine whether a cluster is pure or confounded, we introduce a threshold $\delta \in [0, 1]$. The determination of a cluster can be expressed as:

$$\text{C-type}(C_i) = \begin{cases} \text{Pure}, & \text{if Purity}(C_i) \geq \delta, \\ \text{Confounded}, & \text{if Purity}(C_i) < \delta. \end{cases} \tag{3}$$

Subsequently, the confounder candidates are defined as the set of cluster patterns corresponding to the confounded clusters of the patches in image I. On top of that, the confusing categories $\hat{y}_{conf}$ are defined as the union of categories in confounded clusters. This is formulated as:

$$\mathrm{Confounder}(I_i) = \left\{ \mu(C_k) \,\middle|\, \begin{array}{l} \text{C-type}(C_k) = \text{Confounded}, \\ p_i^j \in C_k, \\ \forall j \in \{1, 2, \ldots, n_i\} \end{array} \right\} \tag{4}$$

$$\hat{y}_{conf}\left(I_i\right) = \bigcup_{\substack{\text{C-type}\ (C_k)=\text{Confounded} \\ p_i^j \in C_k}} \mathcal{Y}\left(C_k\right) \tag{5}$$

where $\mu(C_i)$ is the pattern of cluster C_i, $\mathcal{Y}$ represents the set of categories. Meanwhile, the causal factors for class y are defined as the set of patterns of the pure clusters that are dominated by y:

$$\mathrm{Prototype}_y = \left\{ \mu(C_i) \,\middle|\, \begin{array}{l} \text{C-type}(C_i) = \text{Pure}, \\ p(y \mid C_i) = \max\limits_{y' \in Y} p(y' \mid C_i) \end{array} \right\} \tag{6}$$

Finally, we concatenate the results of Eq. (6) to obtain the dictionary D_C of causal factors.

Stable Features Extracting from Causal Factors

Confounders may arise from co-occurring but non-discriminative objects, as these confounding elements introduce spurious correlations between the image and the label. Front-door intervention is used to cut off the confounding path $P \leftarrow S \rightarrow Y$. Formally, a mediator variable M is introduced to construct a front-door path $P \rightarrow M \rightarrow Y$, and then we indirectly cut off the link $P \rightarrow M$ to block the backdoor path $M \leftarrow P \leftarrow S \rightarrow Y$. Accordingly, the genuine causal effect between P and Y through M is:

$$\begin{aligned} \mathbb{P}(Y|do(P)) = &\sum\nolimits_m \mathbb{P}(M = m|P) \times \\ &\sum\nolimits_{p'} \mathbb{P}(P = p')\mathbb{P}(Y|P = p', M = m) \end{aligned} \tag{7}$$

Following (Chen, Sun, and Zhao 2024; Wang et al. 2024b), we define m as the aggregated feature determined by p: $m = h(p)$, where h is a self-attention module:

$$\begin{aligned} \mathbb{P}(Y|do(P)) &= \sum\nolimits_{p'} \mathbb{P}(Y|M = h(p), P = p')\mathbb{P}(P = p') \\ &= \mathbb{E}_{p'}[\mathbb{P}(Y|m, p')] \end{aligned} \tag{8}$$

Note that $\sum_{p'}$ requires sampling over all patches, which is computationally expensive. Thus, we utilize the dictionary of causal factors to reduce complexity. Specifically, a cross-attention framework is applied to fuse stable causal factors with variable m, formulated as:

$$\mathcal{F}(m, p') = [\mathbb{P}(p') \cdot \mathrm{Softmax}(\frac{(W_q m)^{\mathsf{T}}(W_k D_c)}{\sqrt{d}})](W_v D_C) \tag{9}$$

where W_q,W_K,W_v are linear projections, d is feature dimension, and $\mathbb{P}(p')$ is set to $1/\|D_C\|$. We utilize NWGM (Xu et al. 2015) to absorb the expectation into the forward network and integrate sample-level features with fused features from Eq. (9) to form the intervention prediction:

$$\begin{aligned} \hat{Y}_F = \mathbb{P}(Y|do(P)) &\approx \mathbb{P}(Y|m \oplus \mathcal{F}(m, p')) \\ &\approx \phi(\mathcal{SA}(F_v) \oplus \mathcal{F}(m, p')) \end{aligned} \tag{10}$$

where ϕ is a classifier, $\oplus$ represents features fusion, forming stably discriminative features F_S.

Explicit Confounder Modeling and Debiasing

To explicitly obtain contextually-confusing features F_Z, the EMCM refines sample-specific confounder candidates from the pattern base. Note that the pattern features of confounding clusters, rather than the patch features themselves, serve as confounder candidates. This is motivated by: (1) the pattern features serve as stronger representatives of the visual semantics within a cluster, (2) the reduced number of confounder candidates simplifies training while enhancing the consistency of deconfounding across samples with similar features. The process is formulated as:

$$F_Z = f_{agg}(\mathcal{SA}(\mathrm{Confounder}(I_i))) \tag{11}$$

where confounder features F_z are optimized by the Binary Cross Entropy loss to approximate the predictions of confusing categories, represented by $\mathcal{L}_{BCE}(\phi(F_z), \hat{y}_{conf})$.

	Method	Reference	VireoFood-172		NUS-WIDE			
			acc@1	acc@5	r@1	r@5	p@1	p@5
Visual Causal Methods	ViT-B/16	ICLR'21	88.51	97.66	45.04	86.98	80.18	40.12
	CCD	CVPR'22	88.92	97.84	46.45	<u>89.61</u>	81.69	41.28
	CCIM	CVPR'23	89.18	97.55	46.81	88.37	82.01	40.79
	LGCAM	TPAMI'23	89.23	97.53	46.40	89.27	81.54	41.10
	GOAT	CVPR'24	89.42	97.85	46.92	88.63	82.17	40.91
Cross-modal Alignment Methods	ATNET	MM'19	88.66	94.60	45.59	86.55	80.78	39.89
	FDT	CVPR'23	88.39	96.19	45.59	85.44	80.92	39.30
	IRRA	CVPR'23	89.61	97.97	46.13	86.45	79.96	39.29
	HERM	CVPR'23	<u>90.27</u>	96.88	46.14	85.68	81.31	39.36
	CHAN	CVPR'23	88.38	97.67	46.37	86.94	81.64	39.91
	MOMKE	MM'24	89.77	<u>98.03</u>	46.50	89.14	81.72	41.11
	C2KD	CVPR'24	88.62	97.85	46.19	88.55	81.17	40.84
	MGCC	AAAI'24	88.62	97.96	<u>46.93</u>	89.85	<u>82.25</u>	<u>41.46</u>
	Ours		**92.45**	**98.84**	**48.29**	**90.65**	**84.40**	**41.84**

Table 1: Performance comparison between baselines and EMCM. Our method outperforms the SOTA across all benchmarks. Best and second-best results are bolded and underlined, respectively.

Subsequently, EMCM removes the adverse effects of category-universal contextual information on classification by intervening with $do(S = S')$ and constructing a counterfactual causal graph, as shown in Fig. 3(b). Here, S' essentially represents the contextual semantics shared across different categories in a hypothetical scenario. The counterfactual prediction score Y' is calculated as:

$$\begin{aligned} Y' &= \mathbb{P}(Y|do(S = S'), do(P = P')) \\ &= \mathbb{P}(Y|S = S', P = F_Z) \end{aligned} \tag{12}$$

The debiased global features are defined as the total direct effect of P on Y, i.e., the difference between two predictions given $P = F_G$ and $P = F_Z$. Since F_z captures features shared across categories, the difference $F_G - F_Z$ yields F_C, which are free from confounding contextual information.

Training Strategy

During the forward process, the prediction is calculated as:

$$\hat{Y} = \phi(F_S \oplus F_C) \tag{13}$$

where ϕ is a linear classifier. The model is trained by minimizing the cross-entropy loss $\mathcal{L}_{ce}(\hat{Y}, Y)$. To stabilize model training, the TDE and the front-door intervention branches are trained separately based on their respective features using CE loss. During the training of the TDE branch, $\mathcal{L}_{\text{conf}}$ is integrated into the objective function, weighted by a coefficient α to balance its contribution. Subsequently, the model is fine-tuned based on total loss:

$$\mathcal{L} = \mathcal{L}_{CE}(\hat{Y}, Y) + \alpha \cdot \mathcal{L}_{conf} \tag{14}$$

Experiments

Experiment Settings

Datasets Experiments are conducted on the VireoFood-172 (Chen and Ngo 2016) and NUS-WIDE (Chua et al. 2009) dataset. The former is a single-label dataset of 110,241 food images from 172 classes, while the latter is a multi-label dataset of 269,648 images with 81 concepts.

Evaluation Metrics Following conventional image classification (Meng et al. 2019; Li, Song, and Luo 2017), we leverage accuracy@{1,5} for the VireoFood-172 and precision&recall@{1,5} for the multi-label NUS-WIDE. All reported results are the average of three runs with randomly selected random seeds.

Implementation Details Following standard practice in transformer-based models, we set the feature dimension to 768. The model is trained for 25 epochs with a batch size of 32 using the Adam optimizer. The initial learning rate is selected from the range $[1e{-}5, 1e{-}4]$ with a step size of 0.25, and decayed by a factor of 0.5 every 3 epochs. The DINO threshold is tuned within $[0.1, 0.3]$ with a step of 0.05. The number of clusters K is selected from $\{1024, 2048, 4096\}$ depending on the dataset. All experiments are conducted on four NVIDIA RTX 3090 GPUs.

Performance Comparison

To verify the performance of proposed EMCM, we compare it with various SOTA methods based on causal inference as well as data-driven cross-modal alignment, including CCD (Liu et al. 2022b), CCIM (Yang et al. 2023), LGCAM (Liu, Li, and Lin 2023), GOAT (Wang et al. 2024a), ATNET (Meng et al. 2019), C2KD (Huo et al. 2024), MGCC (Wu et al. 2024a), IRRA (Jiang and Ye 2023), FDT (Chen et al. 2023) HERM (Fu et al. 2023), CHAN (Pan, Wu, and Zhang 2023), MOMKE (Xu, Jiang, and Liang 2024). We implement these methods based on the settings described above for fair comparison. The following observations can be drawn from Table 1:

- The proposed EMCM outperforms SOTA methods on both datasets by a large margin, thanks to the explicit

Modules	VireoFood-172		NUS-WIDE			
	acc@1	acc@5	r@1	r@5	p@1	p@5
Base(B)	88.51	97.66	45.04	86.98	80.18	40.12
B + P	90.45	98.11	46.37	88.81	81.58	40.84
B + P + C	91.23	98.33	47.32	90.04	82.73	41.49
B + P + E	91.46	98.33	47.57	89.66	83.03	41.32
EMCM	**92.45**	**98.82**	**48.30**	**90.68**	**84.44**	**41.86**

Table 2: Ablation study of FACRM. P: Patch Features Fusion. C: Causal Factors Enhancing. E: Explicit Confounder Modeling and Debiasing.

causal modeling of visual elements and complementary causal features intervention.

- Methods that focus on semantically-intensive patches, such as HERM and IRRA, achieve better performance, as the extracted elements have less redundant information than global images. However, they still fall short of EMCM, as they rely solely on data-driven training.
- Certain features alignment methods, due to the lack of information filtering, yield results even worse than the backbone on the challenging VireoFood-172 dataset, which has strong visual diversity.
- Most text-based alignment methods perform moderately on the p@5 metric on the NUS-WIDE dataset, which can be attributed to the noisy nature of the text features in this dataset. Direct alignment leads to low-quality matches between text and images, affecting precision.

Ablation Study

To further validate the contribution of each module of the proposed EMCM, we conducted the ablation experiments presented in Table 2. The following findings are observed:

- The fusion of features extracted from patches significantly outperforms global image classification, especially on the VireoFood-172 dataset, where visual confounding is more severe. This demonstrates that the visual element discretization strategy helps mitigate the impact of noise.
- The stable features derived from causal factors further improve the results of patch features fusion by highlighting class-discriminative semantics and smoothing features for the same category.
- The context debiasing improves global image features by explicitly modeling and removing sample-specific confounder features. Complementary to the causal factors, it focuses on removing the residual contextual bias, thereby forming a more generalized causal representation.

In-depth Analyses

Does the explicit modeling of confounding factors lead to consistent performance gains? Compared with baseline methods such as vanilla patch fusion and the static confounder features (Wang et al. 2024a), the proposed explicit pattern-based modeling with a loss constraint achieves the best overall performance across all metrics. This improvement stems from two key design choices. First, cluster-based confounder modeling outperforms patch selection by using patterns as semantically robust proxies that aggregate

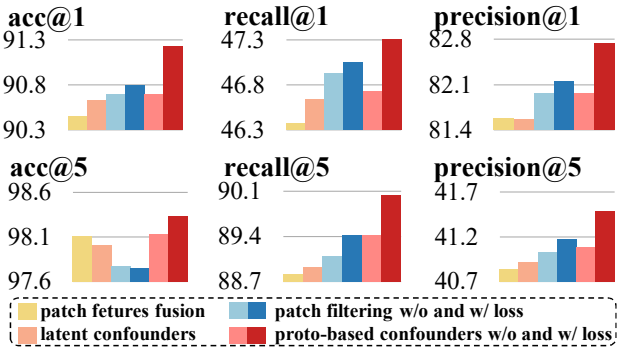

Figure 4: Comparison of variations for confounders. Column 1: VireoFood-172; Column 2-3: NUS-WIDE.

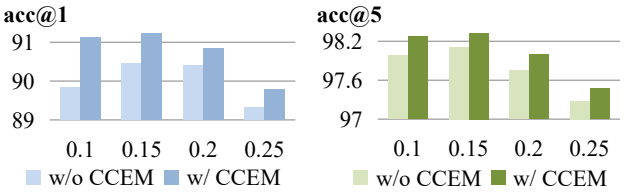

Figure 5: Performance comparison of contextually debiased featurers under different patch qualities.

multiple similar patches. This reduces noise and improves cross-sample consistency. In contrast, patch-level features are often unstable due to semantic ambiguity and inconsistent granularity. Second, the explicit loss constraint provides direct supervision to separate causal and confounding semantics. Apart from that, the low acc@5 performance of patch filtering is likely due to its limited semantic coverage and lack of global structural guidance, which hinders the retrieval of relevant candidate classes.

How does the quality of patches extracted by DINO affect the performance? Patches extracted by the DINO model are associated with confidence scores, which reflect patch quality to some extent. This section evaluates the confounder features derived from the explicit modeling under four confidence thresholds. As shown in Fig. 5, the debiased features consistently outperform vanilla patch feature fusion across all confidence thresholds, demonstrating robustness against varying patch qualities. Notably, at the lowest threshold (0.1), which includes lower-quality patches, it achieves the most significant performance gain, highlighting the ability to filter out redundant information. When the threshold exceeds 0.25, the number of retained patches drops considerably, leading to a general performance decline.

What is the impact of different K values of clustering on EMCM? It is observed that different clustering parameter settings affect the ratio of pure to confounding clusters. To assess their impact on the pattern base and explicit modeling, we analyze how the cluster-patch confusion ratio influences performance under varying K values. As shown in Table 3, increasing K results in more fine-grained and independent clustering of visual features, reducing confusion at the cluster level. However, the patch-level confu-

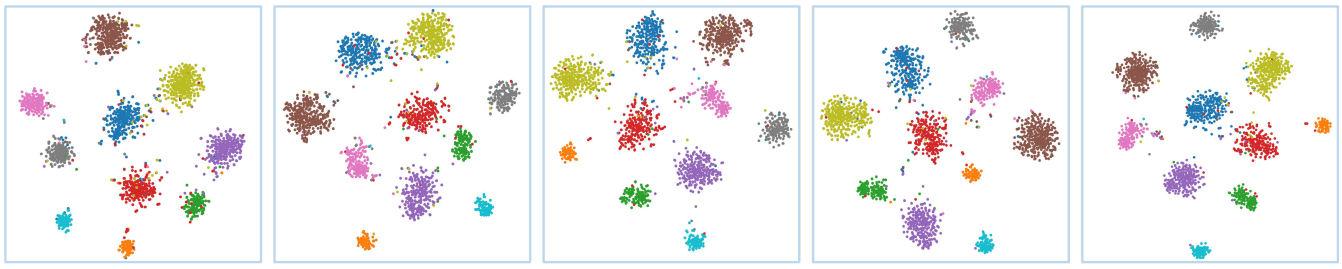

Figure 6: The t-SNE visualization of EMCM modules on 10 classes randomly selected from the VireoFood-172 dataset. (a) features of ViT-B/16, (b) fusion features of patches, (c) features after explicit context debiasing, (d) features after causal factors enhancing, and (e) features of EMCM.

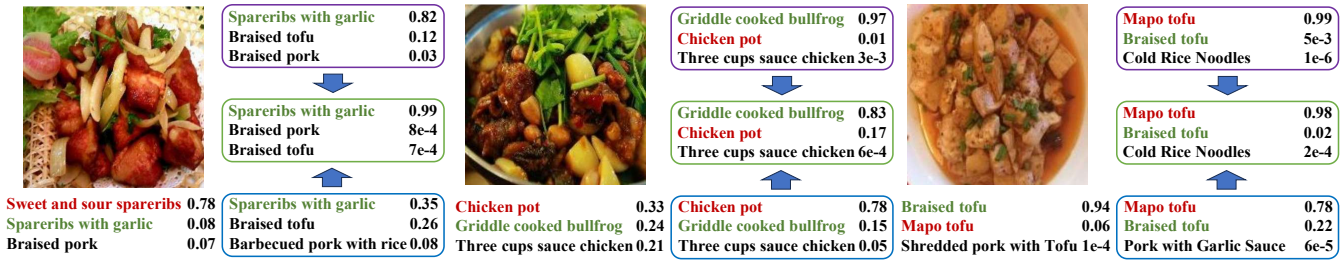

Figure 7: Error analysis of EMCM. The ground-truth labels are shown in green. The blue, purple, and green borders represent the results of context-debiased features, causal factors enhanced features, and EMCM, respectively.

K	$\mathbf{CR}_C$	$\mathbf{CR}_P$		acc@1	acc@5
		train	test		
1024	0.8262	0.3937	0.4020	91.08	98.21
2048	0.7036	0.3910	0.3984	91.23	98.33
4096	0.5742	0.3906	0.3967	91.16	98.23

Table 3: Effect of K in clustering on cluster/patch confusing ratios ($\mathbf{CR}_C$, $\mathbf{CR}_P$) and performance (on VireoFood-172).

sion rate and overall performance remain largely unchanged, suggesting that shared semantic features still contribute to confusion. This demonstrates the stability of the clustering-pattern-based confounder extraction mechanism.

Case Studies

Visualization analysis of the causal and confounding factors. This section explores how sample-specific confounding and causal factors derived from the pattern base contribute to sample representation. t-SNE visualizations of 10 random classes from the VireoFood-172 dataset are shown in Fig. 6. Initially, significant overlap is observed (Fig. 6(a)). After patch extraction and feature fusion (Fig. 6(b)), some category confusion is reduced, but cluster divergence increases. The explicit context debiasing (Fig. 6(c)) further improves class separation by removing unstable features, while causal factors (Fig. 6(d)) enhance inter-class differences. Finally, EMCM (Fig. 6(e)) results in well-separated features with minimal overlap, demonstrating the effectiveness of EMCM in disentangling discriminative causal semantic features and confounding features across multiple categories.

Error Analysis. This section assesses EMCM's generalization through successful and failed cases, as shown in Fig. 7. In the left sample, ECMD significantly lowers the scores of the confusing categories, and causal factors-enhanced features increase the score of the correct category. The middle sample suggests that, for instances where raw ingredients are atypical, merely removing interference from ingredients like parsley leads to increased confusion. Yet by introducing front-door interventions, the model progresses in handling unobservable semantic bias. The right sample demonstrates that excessive intervention by EMCM can cause new errors in confused categories. This occurs because the model attenuates attention on the shared features of tofu between mapo tofu and braised tofu, instead placing more emphasis on the soup base, which is more typical in mapo tofu.

Conclusion

To address the limitations of latent confounder modeling in capturing object-level spurious correlations under complex scenes, EMCM explicitly extracts causal and confounding factors from visual semantic patterns and fuses causal stable features with debiased contextual information for generalized representations. A pre-trained DINO model is used to extract visual patches, and clustering is applied to derive semantic patterns. Based on cluster purity measurement, causal and confounding factors are separated. Furthermore, the fusion of causal stable features and debiased context features leads to improved generalization capability. Future work will fully leverage the hierarchical relationships between image patches and their associated textual words for more fine-grained semantic disentanglement. Moreover, constructing a cross-modal causal graph that connects visual elements and semantic words to enable more structured causal reasoning remains an open challenge.

References

Andonian, A.; Chen, S.; and Hamid, R. 2022. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16430–16441.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, J.; and Ngo, C.-W. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM international conference on Multimedia*, 32–41.

Chen, K.; Sun, S.; and Zhao, J. 2024. Camil: Causal multiple instance learning for whole slide image classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1120–1128.

Chen, W.; Liu, Y.; Wang, C.; Zhu, J.; Zhao, S.; Li, G.; Liu, C.-L.; and Lin, L. 2024. Cross-Modal Causal Intervention for Medical Report Generation. arXiv:2303.09117.

Chen, Y.; Yuan, J.; Tian, Y.; Geng, S.; Li, X.; Zhou, D.; Metaxas, D. N.; and Yang, H. 2023. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15095–15104.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, 1–9.

Fu, Z.; Mao, Z.; Song, Y.; and Zhang, Y. 2023. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15159–15168.

Gao, Y.; Liu, J.; Xu, Z.; Wu, T.; Zhang, E.; Li, K.; Yang, J.; Liu, W.; and Sun, X. 2024. Softclip: Softer cross-modal alignment makes clip stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1860–1868.

Guan, Q.-L.; Zheng, Y.; Meng, L.; Dong, L.-Q.; and Hao, Q. 2023. Improving the generalization of visual classification models across IoT cameras via cross-modal inference and fusion. *IEEE Internet of Things Journal*.

Huo, F.; Xu, W.; Guo, J.; Wang, H.; and Guo, S. 2024. C2KD: Bridging the Modality Gap for Cross-Modal Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16006–16015.

Jiang, D.; and Ye, M. 2023. Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2787–2797.

Li, W.; Su, X.; Song, D.; Wang, L.; Zhang, K.; and Liu, A.-A. 2023. Towards Deconfounded Image-Text Matching with Causal Inference. In *Proceedings of the 31st ACM International Conference on Multimedia*, 6264–6273.

Li, Y.; Song, Y.; and Luo, J. 2017. Improving pairwise ranking for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3617–3625.

Liu, B.; Wang, D.; Yang, X.; Zhou, Y.; Yao, R.; Shao, Z.; and Zhao, J. 2022a. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18041–18050.

Liu, R.; Liu, H.; Li, G.; Hou, H.; Yu, T.; and Yang, T. 2022b. Contextual debiasing for visual recognition with causal mechanisms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12755–12765.

Liu, Y.; Li, G.; and Lin, L. 2023. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Y.; Qin, G.; Chen, H.; Cheng, Z.; and Yang, X. 2024. Causality-Inspired Invariant Representation Learning for Text-Based Person Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14052–14060.

Meng, L.; Chen, L.; Yang, X.; Tao, D.; Zhang, H.; Miao, C.; and Chua, T.-S. 2019. Learning using privileged information for food recognition. In *Proceedings of the 27th ACM international conference on multimedia*, 557–565.

Pan, Z.; Wu, F.; and Zhang, B. 2023. Fine-grained image-text matching by cross-modal hard aligning network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19275–19284.

Pearl, J.; Glymour, M.; and Jewell, N. P. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Wang, L.; He, Z.; Dang, R.; Shen, M.; Liu, C.; and Chen, Q. 2024a. Vision-and-Language Navigation via Causal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13139–13150.

Wang, Y.; Meng, L.; Ma, H.; Wang, Y.; Huang, H.; and Meng, X. 2024b. Modeling Event-level Causal Representation for Video Classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3936–3944.

Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; and Wu, F. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10941–10950.

Wu, X.; Ma, W.; Guo, D.; Zhou, T.; Zhao, S.; and Cai, Z. 2024a. Text-based Occluded Person Re-identification via Multi-Granularity Contrastive Consistency Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6162–6170.

Wu, Y.; Liu, Y.; Zhao, Z.; Lu, W.; Zhang, Y.; Sun, C.; Wu, F.; and Kuang, K. 2024b. De-biased Attention Supervision

for Text Classification with Causality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19279–19287.

Xie, C.-W.; Wu, J.; Zheng, Y.; Pan, P.; and Hua, X.-S. 2022. Token embeddings alignment for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4555–4563.

Xu, K.; Ba, J. L.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, 2048–2057.

Xu, W.; Jiang, H.; and Liang, X. 2024. Leveraging Knowledge of Modality Experts for Incomplete Multimodal Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 438–446.

Yang, D.; Chen, Z.; Wang, Y.; Wang, S.; Li, M.; Liu, S.; Zhao, X.; Huang, S.; Dong, Z.; Zhai, P.; et al. 2023. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19005–19015.

Yang, J.; Duan, J.; Tran, S.; Xu, Y.; Chanda, S.; Chen, L.; Zeng, B.; Chilimbi, T.; and Huang, J. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15671–15680.

Yang, X.; Zhang, H.; Qi, G.; and Cai, J. 2021. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9847–9857.

Zang, C.; Wang, H.; Pei, M.; and Liang, W. 2023. Discovering the real association: Multimodal causal reasoning in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19027–19036.

Zhang, H.; Zhang, L.; Zhang, K.; and Mao, Z. 2024. Identification of Necessary Semantic Undertakers in the Causal View for Image-Text Matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7105–7114.

Zhang, X.; Zhang, F.; and Xu, C. 2023. Reducing Vision-Answer biases for Multiple-choice VQA. *IEEE Transactions on Image Processing*.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) yes

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) partial

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) no

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) partial

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes