Out-of-Context Misinformation Detection via Variational Domain-Invariant Learning with Test-Time Training

Xi Yang[1], Han Zhang[1], Zhijian Lin[1], Yibiao Hu[1], Hong Han [1*]

[1] School of Electronic Engineering, Xidian University, Shaanxi, China

yangxi@stu.xidian.edu.cn, 22021110280@stu.xidian.edu.cn, 23021211252@stu.xidian.edu.cn, yibiaohu@stu.xidian.edu.cn, hanh@mail.xidian.edu.cn

Abstract

Out-of-context misinformation (OOC) is a low-cost form of misinformation in news reports, which refers to place authentic images into out-of-context or fabricated image-text pairings. This problem has attracted significant attention from researchers in recent years. Current methods focus on assessing image-text consistency or generating explanations. However, these approaches assume that the training and test data are drawn from the same distribution. When encountering novel news domains, models tend to perform poorly due to the lack of prior knowledge. To address this challenge, we propose **VDT** to enhance the domain adaptation capability for OOC misinformation detection by learning domain-invariant features and test-time training mechanisms. Domain-Invariant Variational Align module is employed to jointly encodes source and target domain data to learn a separable distributional space domain-invariant features. For preserving semantic integrity, we utilize domain consistency constraint module to reconstruct the source and target domain latent distribution. During testing phase, we adopt the test-time training strategy and confidence-variance filtering module to dynamically updating the VAE encoder and classifier, facilitating the model's adaptation to the target domain distribution. Extensive experiments conducted on the benchmark dataset NewsCLIPpings demonstrate that our method outperforms state-of-the-art baselines under most domain adaptation settings.

Code — https://github.com/yanggxii/VDT
Extended version — https://arxiv.org/pdf/2511.10213

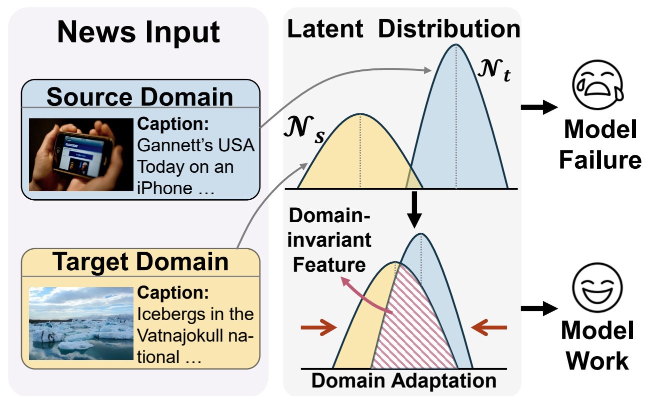

Figure 1: The illustration of domain adaptation challenge in out-of-context misinformation detection. Model failure on target domain due to the significant distribution gap between the source and target domains. By domain adaptation to learn domain-invariant features, the model adapted to target domain better.

Introduction

The news reports on online platform is numerous and various. Particularly, the rapid advancement of generative AI technologies has led to a rapid rise in synthetic data (Pan et al. 2023; Chen and Shu 2023), which presents a significant challenge to the authenticity and credibility of news reports. In contrast to Deepfake model (Dolhansky et al. 2020) which generates some non-existent images, a low-cost and low-tech type of misinformation is reusing of images (Jaiswal et al. 2017; Fazio 2020; Qi et al. 2024). This misinformation places the real picture in a out-of-context or a false image-text pair environment, often referred to as out-of-context misinformation(OOC) (Hu et al. 2024). Due to its low technical requirements, misinformation is easy to manipulate, resulting in widely spreading and hard detecting.

Faced with this challenge, current research can be divided into two directions roughly. Some studies tend to calculate the similarity of pictures and text in the representation space (Jaiswal et al. 2017; Luo, Darrell, and Rohrbach 2021; Huang et al. 2022) or use external evidence to provide support (Müller-Budack et al. 2020; Papadopoulos et al. 2023, 2025). While other works places greater emphasis on the interpretative side of detection (Qi et al. 2024; Zhang et al. 2023b). These approaches leverage Multimodal large language models (MLLMs) to generate plausible and well-grounded explanations that justify the detection outcomes. Although some progress has been made in these methods, they all assume that the training and testing data are drawn from the same distribution. In general, news reports encompass a diverse array of topics and domains, with significant variations in thematic style and content across different subjects. Due to lack of prior knowledge, poor performance of

[*]Corresponding author

detection models will occur when new news topics emerge. It is not feasible to retrain the detection model every time a new domain arises, and it is equally difficult to produce datasets for each new topic. Figure 1 illustrates the domain adaptation challenge in ooc misinformation detection.

Domain adaptation is an important research direction in transfer learning (Lu, Tong, and Ye 2025), aiming to address the fundamental challenge of distribution discrepancy between source and target domain. Specifically, it enhances model performance on the target domain by utilizing the rich labeled data available in the source domain. Many studies (Ganin et al. 2016; Zhang et al. 2019; Wang et al. 2022; Tzeng et al. 2014) have demonstrated that learning domain-invariant features can effectively mitigate the distribution shift between source and target domains. These features effectively capture the essential information of data in both source and target domain, while remaining robustness to domain-specific attributes and label distributions. However, in the out-of-context news detection task, there are still some challenges. First, OOC datasets are limited in scope, covering only a few topics and institutions. The substantial difference in data characteristics across different domains poses significant challenges for models to learn generalizable domain-invariant features. Second, the inherent lack of contextual information in manipulated data, which often leads models to over-rely on certain local multimodal features (Geirhos et al. 2020; Zhang et al. 2021), thereby exacerbating the difficulty of learning robust domain-invariant representations.

To deal with this challenge, we proposed **V**ariational **D**omain-Invariant Learning with **T**est-Time Training (**VDT**) for domain adaptive out-of-context news detection. VDT first uses an Multimodal Large Language Model (MLLM) to directly encode the image and text into a multimodal feature representation. And then we introduce a domain-invariant variational alignment module that employs a shared VAE to jointly encode the source and target data, learn their latent distributions, and promote feature alignment across domains. Moreover, we leverage the properties of the mean and variance to enhance the representational capacity of the domain-invariant features. To reduce the domain gap and ensure that domain-invariant features are captures, we impose constraints on the latent distributions between the source and target domains. Meanwhile, we employ a domain consistency constraint module to reconstruct the latent distributions to prevent representation collapse. Finally, we adopt test-time training strategy to enhance the model's adaptability to the target domain. During the evaluation phase, a confidence-variance filtering mechanism is applied to screen pseudo-labeled (Lee et al. 2013) that from unseen target domain. High-quality samples are retained to update the VAE encoder and the classifier.

The main contributions of our work are four-fold:

- We propose a variational domain-invariant learning with test-time training framework to address the performance degradation caused by distribution shifts between the source and target domains in out-of-context misinformation detection, thereby improving the model's domain adaptation capability.
- To prevent the model from relying on multimodal local features due to inherent lack of contextual information in manipulated data, we introduce the DIVA module to learn latent distributions. By imposing constraints to reduce the domain gaps, ensuring that domain-invariant features are effectively captured.
- During the test phase, we adopt a test-time training strategy to dynamically update the VAE encoder and classifier. confidence-variance filtering mechanism is employed to select high-quality samples, thereby model can better adapt to the target domain.
- Extensive experiments on the widely-used benchmark datasets, NewsCLIPpings, demonstrate that our proposed method achieves strong and competitive performance under various domain adaptation settings.

Related Work

Domain Adaptation News Detection

Current research primarily focuses on evaluating image-text consistency (Jaiswal et al. 2017; Aneja, Bregler, and Nießner 2021; Luo, Darrell, and Rohrbach 2021; Mu, Das Bhattacharjee, and Yuan 2023; Papadopoulos et al. 2023; Yuan et al. 2023) and leveraging large language models to enhance interpretability (Papadopoulos et al. 2023; Zhang et al. 2023a; Papadopoulos et al. 2025). But the domain adaptation challenges to out-of-context misinformation detection is neglected. Many studies have adopted adversarial learning (Wang et al. 2018; Yuan et al. 2021; Zhang et al. 2020) to extract domain-invariant features.

Alternative research directions have explored discrepancy-based approaches (Yue et al. 2022; Gu et al. 2024), which achieve domain adaptation by minimizing the distribution divergence between source and target domains.

Test-Time Training

Test-Time Training methods (Wu et al. 2024) update the model at inference time using a Y-shaped architecture, in which the main task and a self-supervised auxiliary task are jointly optimized during training.

Sun et al. (Osowiechi et al. 2023; Sun et al. 2020) introduced a seminal technique that later gave rise to a family of methods collectively known as Test-Time Training (TTT), which enables re-training the network on the test set even though no test labels are available. The original work employed rotation prediction (Komodakis and Gidaris 2018) as the auxiliary task. Later approaches (Gandelsman et al. 2022; Liu et al. 2021b) replaced the original auxiliary task with either Masked Autoencoder reconstruction (He et al. 2022) or contrastive learning (Chen et al. 2020).

In our work, we adopt VAE reconstruction as the auxiliary task, and update the VAE encoder during the TTT phase. The decoder is kept frozen to prevent it from being influenced by target domain samples, which may distort the original reconstruction structure. Moreover, since the target domain data are not seen by the classifier during training, the classifier may exhibit a source-domain bias. To mitigate this, we also update the classifier parameters during test-time training.

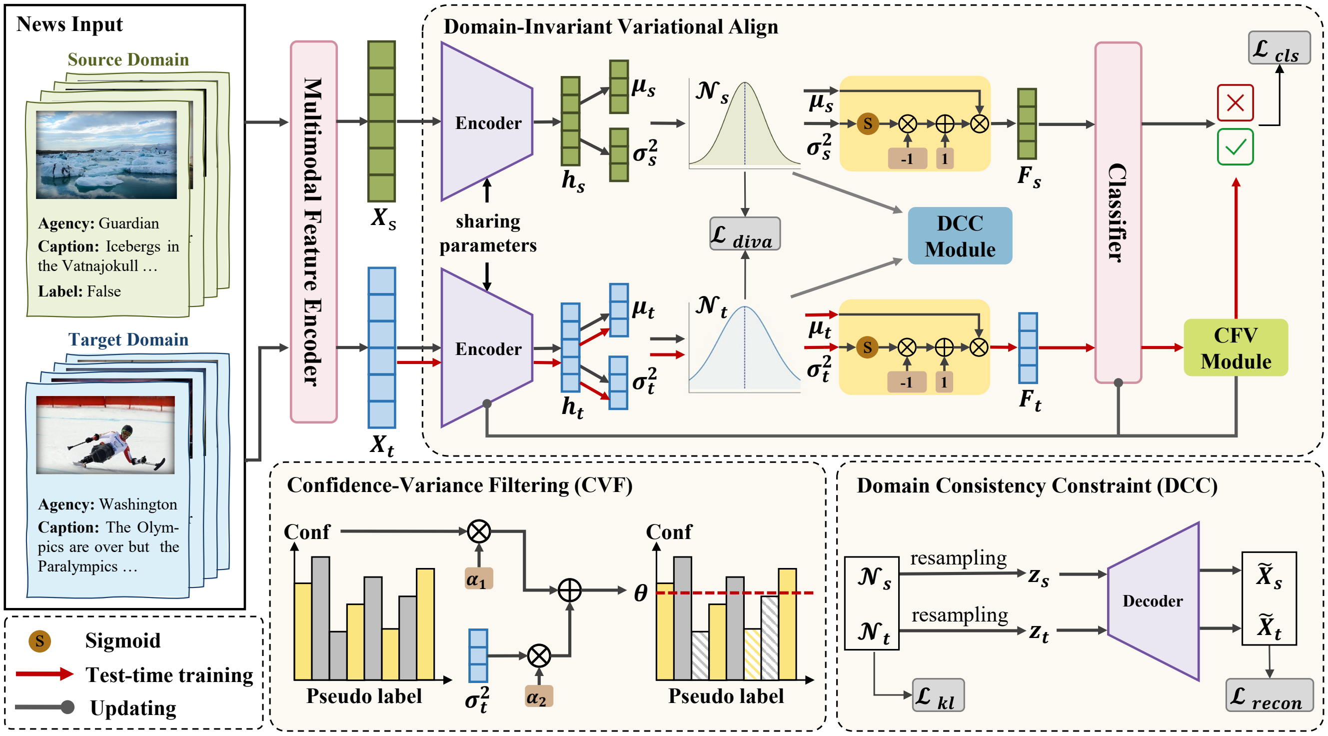

Figure 2: The illustration of proposed VDT model.

Method

Multimodal Feature Encoder

Following Gu et al. (Gu et al. 2024), we use MLLM to encode the news text and image into a unified semantic representation space, for capturing the semantic differences between original news and out-of-context news. It enables the model to learn rich cross-modal interactions, facilitating more accurate detection of semantic inconsistencies. Formally, the multimodal feature encoder is defined as follows:

$$\begin{aligned} X^s &= MLLM((x_{img}, x_{txt})^s) \\ X^t &= MLLM((x_{img}, x_{txt})^t) \end{aligned} \tag{1}$$

where X^S and X^T denote the multimodal features representation of the news image-text pair (x_{img}, x_{txt}) from the source and target domain.

Domain-Invariant Variational Align

After encoding the image-text pairs in news articles, we introduce a Domain-Invariant Variational Align (DIVA) module to model the latent distributions of source and target domain samples.

In DIVA module, we adopt Variational Autoencoder (VAE) (Kingma and Welling 2013; Park et al. 2024), whose generative capacity facilitates the learning of separable latent features for both the source and target domain distributions. Specifically, the latent distributions of source and target domains are learned through two VAE encoders that share parameters. The process is defined as follows:

$$h_s = E_s(X_s; \theta^e) \qquad h_t = E_t(X_t; \theta^e) \tag{2}$$

where θ^e is denotes the learnable parameters of the encoder.

The latent representations of the source h_S and target domains h_t are further passed through distinct linear layers to obtain the mean and variance respectively, as follows:

$$\mu_s = W_{s\mu} h_s + b_{s\mu} \qquad log\sigma_s^2 = W_{s\sigma} h_s + b_{s\sigma} \tag{3}$$

where μ_s is the mean of the latent distribution for the source domain, $log\sigma_s^2$ represent the variance. $W_{s\mu}$, $b_{s\mu}$ denote the weights and bias of the mean layer, and $W_{s\sigma}$, $b_{s\sigma}$ represent the weights and bias of the log-variance layer.

Similarly, the mean and variance of the target domain can be obtained in the same manner, represent as μ_t and σ_t^2. In this way, we obtain Gaussian distributions in the latent space for both the source and target domains, denoted as: $\mathcal{N}_s \sim (\mu_s, \sigma_s^2)$ and $\mathcal{N}_t \sim (\mu_t, \sigma_t^2)$.

The mean μ represents the semantic center of samples in the latent space, while the variance σ^2 controls the degree of dispersion. A larger variance indicates that the sample is more spread out, farther from the semantic center, and associated with higher uncertainty.

To construct more robust domain-invariant features, we transform the variance into gating weights and express the final domain-invariant feature as:

$$F_s = \mu_s(1 - gate_s) \qquad gate_s = sigmoid(log\sigma_s^2) \tag{4}$$

This representation ensures that samples with high uncertainty are attenuated in weight, effectively reducing the influence of features far from the semantic center. Conversely, samples with low variance retain their weights.

Contrastive learning has been shown to effectively pull semantically similar samples closer while pushing dissimilar ones apart in the latent space (Bhattacharjee et al. 2023). We adopt a contrastive loss to constrain the mean representations of the source and target domains, aiming to align their distributions and reduce the domain gap. We refer to this objective as the DIVA loss, which is formulated as follows:

$$\mathcal{L}_{diva} = -log \sum_{i \in N} \frac{exp(sim(\mu_{si}, \mu_{ti})/\tau)}{\sum_{j=1, j\neq i}^{2N} exp(sim(\mu_{sj}, \mu_{tj})/\tau)} \quad (5)$$

where N denotes the batch size, and μ_{si} and μ_{ti} represent the mean vectors of the i-th sample from the source and target domains, respectively. τ is a temperature parameter, and $sim(\cdot)$ denotes the similarity metric, for which we use cosine similarity.

Domain Consistency Constraint Module

After obtaining the latent distributions of the source and target domains, we employ the reparameterization trick to enable gradient-based optimization.

$$\begin{aligned} z_s &= \mu_s + \sigma_s \cdot \epsilon, \quad \epsilon \in \mathcal{N}(0, I) \\ z_t &= \mu_t + \sigma_t \cdot \epsilon, \quad \epsilon \in \mathcal{N}(0, I) \end{aligned} \quad (6)$$

where random noise ϵ is sampled from a normal distribution $\mathcal{N}(0, I)$.

To prevent the encoder from losing critical information while aligning the source and target domains—i.e., to avoid semantic collapse caused by over-alignment. We pass the sampled latent codes z_s and z_t through a shared decoder to reconstruct the original inputs. Formally, this can be expressed as:

$$\hat{X}_s = D_s(z_s; \theta^d) \qquad \hat{X}_t = D_t(z_t; \theta^d) \quad (7)$$

where θ^d is denotes the learnable parameters of the decoder.

We formulate the reconstruction loss ensuring the distributions effectively capture critical semantic features and maintain semantic fidelity, as follows:

$$\mathcal{L}_{recon} = ||X_s - \hat{X}_s||^2 + ||X_t - \hat{X}_t||^2 \quad (8)$$

As training epochs number increases, the encoder progressively compresses the data into an extremely narrow region of the latent space, diminished expressiveness of the latent variables. To mitigate such representational degradation, we introduce a Kullback–Leibler (KL) divergence (Hershey and Olsen 2007) loss as a regularization term to constrain the latent distributions of both the source and target domains toward a standard normal distribution $\mathcal{N}(0, I)$. This way encourages the target domain representations to be implicitly aligned with those of the source domain, thereby promoting domain alignment and improving the coherence and sampleability of the learned latent space. The KL divergence loss is defined as:

$$\begin{aligned} \mathcal{L}_{kl} &= \frac{1}{2} \sum_{m \in \{s,t\}} KL(\mathcal{N}(\mu_m, \sigma_m^2) \,||\, \mathcal{N}(0, I)) \\ &= \frac{1}{2} \sum_{m \in \{s,t\}} \frac{1}{2}(\mu_m^2 + \sigma_m^2 - log\sigma_m^2 - 1) \end{aligned} \quad (9)$$

Therefore, the overall loss of domain consistency constraint module can be expressed as:

$$\mathcal{L}_{dcc} = \mathcal{L}_{recon} + \beta \mathcal{L}_{kl} \quad (10)$$

where the β parameter is used to weight the Kullback-Leibler (KL) divergence term in the Evidence Lower Bound (ELBO) objective (Burgess et al. 2018). It serves as a critical hyperparameter that balances reconstruction fidelity and latent disentanglement in the VAE framework.

Training

The final domain-invariant feature F is used for classification to obtain the predicted labels of out-of-context news instances, formulated as:

$$\hat{y} = \arg\max_c(f) \qquad f = CLS(F) \quad (11)$$

The loss function for the classifier is the cross-entropy loss, where $\hat{y}$ is the predicted label and y is the ground truth label:

$$\mathcal{L}_{cls} = -E_{y \sim Y}[ylog(\hat{y}) + (1-y)log(1-\hat{y})] \quad (12)$$

We combine classification loss $\mathcal{L}_{cls}$, the domain-invariant variational align loss $\mathcal{L}_{diva}$, and the domain consistency constraint loss $\mathcal{L}_{dcc}$ as optimization objective. The overall loss is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cls} + \lambda_2 \mathcal{L}_{diva} + \lambda_3 \mathcal{L}_{dcc} \quad (13)$$

Test-Time Training

To further adapt the model to the target domain, we introduce a test-time training strategy that leverages the pretrained network to adjust its parameters using the unlabeled target domain.

First, we freeze all model parameters except for the VAE encoder and the classifier. Then, the DIVA module is utilized to obtain the mean and variance of the target domain test samples, from which the domain-invariant feature F is derived, formulated as:

$$F_t = \mu_t(1 - sigmoid(log\sigma_t^2)) \quad (14)$$

The pseudo-labels and confidence scores are obtained through the classifier as follows:

$$\widetilde{y} = \arg\max_c(f_t) \qquad f_t = CLS(F_t) \quad (15)$$

$$conf = \max_c(f) \quad (16)$$

The confidence score conf corresponds to the maximum value among the two-dimensional predicted probabilities, indicating the model's degree of certainty in its current prediction.

Confidence-Variance Filtering Module To further improve the quality of pseudo-labels, we design a Confidence-Variance Filtering (CVF) Mechanism.

At this stage, we have access to the variance features from the representation layer and the confidence scores for the unlabeled target domain samples. The variance reflects the representational reliability of a sample—whether its features

Agency	train		test	
	class0	class1	class0	class1
Guardian	283782	283230	29631	29532
BBC	108060	97908	11561	10513
Washington Post	89914	91613	9217	9471
USA Today	109694	118699	11675	12568

Table 1: The statistics of NewsCLIPpings dataset. The four quantities in each agency are respectively train set-class0, train Set-class1, test set-class0, and test set-class1.

are stable and can effectively characterize its semantics. The confidence score indicates the predictive reliability, that is, the degree of certainty of the model's classification decision.

Only when the samples simultaneously meet the conditions of "high confidence + low variance" are selected as high-quality pseudo-label samples.

This mechanism can be specifically expressed as:

$$score = \alpha_1 \cdot conf + \alpha_2 \cdot \sigma_t^2 \tag{17}$$

And set a threshold θ. We only retain the data with scores higher than the threshold θ to update the VAE encoder and classifier.

Updating The filtered data is reconstructed to obtain $\widetilde{X}_t$. During test-time training, the loss function consist of the cross-entropy loss, reconstruction loss, and KL divergence loss, which is expressed as:

$$\mathcal{L}_{TTT} = \mathcal{L}_{cls} + \mathcal{L}_{dcc} \tag{18}$$

Experiments

Dataset and Evaluation Metrics

Dataset We conduct experiments on NewsCLIPpings dataset (Luo, Darrell, and Rohrbach 2021), which is collected from VisualNews dataset (Liu et al. 2021a), a benchmark news image caption dataset. The sources of the news include four institutions: Guardian (G), BBC (B), Washington Post (W) and USA Today (U). The labels are balanced, and the distribution of samples across different news agencies is detailed in Table 1. In this dataset, we treat different news agencies as distinct domains to address the domain adaptation problem for OOC misinformation detection.

Evaluation Metrics For evaluation metrics, we report accuracy and macro $F1$ score. Alongside the main metrics, we further include $F1_{real}$ and $F1_{fake}$ (Hu et al. 2024) to better understand class-wise performance.

Implementation details

The proposed VDT model is implemented in PyTorch 1.13.1 with CUDA 11.6, and all experiments are run on a single NVIDIA RTX 3080 Ti GPU. And We use BLIP-2's multimodal feature extractor (Li et al. 2023) to embed the image and text, and obtain the embedding x of size 768. We employ the Adam optimizer for training with a batch size of 256 for 20 epochs. The early stopping epoch is set to 5. In training stage, the learning rate is set to $1\mathrm{e}-4$ and in test-time training phase, the learning rate is set to $2\mathrm{e}-5$. The dim of μ and σ^2 is set to 128, and the dim of classifier layer is set to 500.

In the loss function, the weight of kl divergence loss β is set to 1.5. λ_1, λ_2, and λ_3 is set to 2, 0.5 and 1. The α_1 is set to 2 and the α_2 is set to -1. The threshold θ is set to 0.9.

For NewsCLIPpings dataset, inspired by Gu et al. (Gu et al. 2024), we respectively used two news organizations from the same country as the target domain and the remaining two news organizations as the source domain. During the training phase, we have access to labeled source domain data and unlabeled target domain data. In the testing phase, we evaluate the model's performance on the labeled target domain data.

Compare to the State-Of-The-Art

We conduct a fair and reasonable comparison between VDT and six representative, republic state-of-the-art(SOTA) methods, including: EANN (Wang et al. 2018), MDA-WS (Li et al. 2021), CANMD (Yue et al. 2022), REAL-FND (Mosallanezhad et al. 2022), CADA (Mosallanezhad et al. 2022), and ConDA-TTT (Gu et al. 2024). On the NewsCLIPpings dataset, our proposed VDT outperforms baseline models under the vast majority of domain adaptation settings. Specifically, when targeting domain G (Guardia), VDT achieves improvements of 1.19% in $F1$ score and 0.97% in accuracy compared to ConDA-TTT. Similarly, when domain U (USA Today) is used as the target, our method surpasses the strongest baseline by 0.93% in $F1$ score and 0.61% in accuracy.

Experimental results demonstrate that our proposed model effectively aligns the distributions of the source and target domains, which reduces the domain gap and enables the extraction of discriminative domain-invariant features, thereby enhancing the model's generalization performance on the target domain.

Different Domain Adaptation Setting

To complement the main experiments and cover domain adaptation settings not previously included, we conduct additional experiments to further validate the effectiveness of the proposed method in domain adaptation tasks. The results are shown in Table 3.

Table 3 includes three scenes: using a single domain, three domains, and all domains as the source domain. The experimental results show that in the vast majority of domain adaptation settings, our proposed method consistently outperforms the state-of-the-art model ConDA-TTT. In particular, for the BGW→U setting, our method outperforms ConDA-TTT by 2.36% in $F1$ score and 2.03% in accuracy. Significant performance differences between B→U & U→B and G→U & U→G indicate a strong asymmetry between these domains. The adaptive direction has a notable impact on performance (Farahani et al. 2021; You et al. 2019), likely due to the higher diversity or quality of source domains (e.g., G), which makes them more effective as source domains. Conversely, target domains like U may have simpler data distributions.

Furthermore, we observe that multi-source adaptation generally performs better than single-source adaptation. For

Model	U, W → B				U, W → G				B, G → U				B, G → W			
	$F1$	Acc	F_{real}	F_{fake}	$F1$	Acc	F_{real}	F_{fake}	$F1$	Acc	F_{real}	F_{fake}	$F1$	Acc	F_{real}	F_{fake}
EANN	68.72	69.82	65.18	72.26	75.93	76.05	74.30	77.56	79.80	79.70	78.06	81.53	78.33	78.33	77.07	79.58
MDA-WS	70.02	69.99	72.20	67.85	74.95	75.23	72.39	77.52	79.51	79.46	77.64	81.39	78.14	78.14	76.92	79.36
CANMD	68.61	68.53	66.70	71.37	74.78	75.06	74.01	77.79	80.30	80.29	78.67	81.85	78.60	78.50	77.40	78.62
REAL-FND	69.61	69.84	66.81	72.42	75.39	75.59	73.14	77.63	80.33	80.45	78.78	81.87	78.75	78.77	78.10	79.41
CADA	68.73	69.85	65.12	72.34	75.65	75.83	73.66	77.64	80.31	80.26	78.83	81.80	78.27	78.27	77.24	79.31
ConDA-TTT	71.86	71.83	73.78	69.94	75.40	75.66	78.02	72.79	81.03	81.34	81.87	80.18	78.61	78.74	79.62	77.61
VDT	**72.26**	**72.30**	74.08	70.26	**76.59**	**76.63**	77.53	75.65	**81.96**	**81.95**	81.38	82.49	**79.02**	**79.02**	78.61	79.42

Table 2: Performance (%) comparison between VDT and baselines on NewsCLIPpings. B, G, U and W stand for BBC, Guardian, USA Today and Washington Post respectively. In X → Y, X denotes the source domain, Y denotes the target domain. The best performance is in bold text. The second best performance is underlined.

Mode	ConDA-TTT		VDT	
	F1	Acc	F1	Acc
B→U	77.16	77.63	75.96	76.04
U→B	71.23	71.19	71.55	71.60
B→W	76.68	76.92	76.84	76.86
W→B	72.01	71.99	71.75	71.75
B→G	73.59	74.16	74.55	74.56
G→B	73.27	73.23	72.63	72.62
G→U	81.34	81.59	81.65	81.66
U→G	75.07	75.31	73.95	74.40
G→W	79.00	79.11	79.12	79.12
W→G	74.73	74.98	75.79	75.96
U→W	79.30	79.33	79.27	79.26
W→U	82.27	82.49	83.28	83.27
GUW→B	72.18	72.15	71.85	71.84
BUW→G	75.33	75.45	77.14	77.15
BGW→U	80.42	80.74	82.78	82.77
BGU→W	78.52	78.60	79.76	79.78
ALL→B	74.01	73.96	74.72	74.86
ALL→G	76.64	76.81	78.21	78.21
ALL→U	82.77	83.04	84.52	84.51
ALL→W	80.18	80.31	80.73	80.74

Table 3: Performance (%) of different domain adaptation settings on NewsCLIPpings. In X → Y, X denotes the source domain, Y denotes the target domain. "ALL" include four agencies (B, G, U, and W).

		VDT	$w/o\,\mathcal{L}_{diva}$	$w/o\,\mathcal{L}_{dcc}$	w/o TTT	w/o CVF
B	$F1$	72.26	68.87	62.80	69.08	68.13
	Acc	72.30	69.12	63.87	69.25	67.77
G	$F1$	76.59	73.71	68.41	75.30	74.27
	Acc	76.63	74.29	69.36	75.43	74.57
U	$F1$	81.96	79.82	72.06	79.57	79.84
	Acc	81.95	79.90	72.27	79.72	79.85
W	$F1$	79.02	78.23	70.87	77.98	77.27
	Acc	79.02	78.30	71.86	77.99	77.43

Table 4: Ablation study on NewsCLIPpings dataset. w/o indicates the ablation of the corresponding component. B, G, U, and W denote the target domains, and the experimental settings are consistent with those in Table 2.

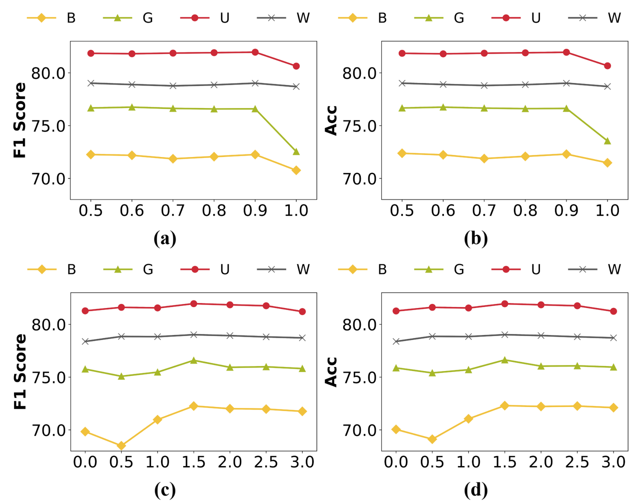

Figure 3: VDT's performances with different β (sub-figure (a) and (b)) and threshold θ (sub-figure (c) and (d)) values. The legend illustrates target domains.

example, ALL→U achieves better performance than G→U and W→U. Our proposed method performs particularly well in multi-source settings, demonstrating its superior effectiveness and generalization capability.

Ablation Study

We conduct an ablation study on different components of VDT to verify the effectiveness of each proposed module, and the results are shown in Table 4. Specifically, we ablate $\mathcal{L}diva$, $\mathcal{L}dcc$, the CVF mechanism, and TTT.

From the experimental results, we observe that removing any of these components leads to varying degrees of performance degradation. In particular, ablating $\mathcal{L}dcc$ causes the most significant drop in performance. This is mainly because $\mathcal{L}dcc$ is responsible for the VAE reconstruction task. Without it, the latent space lacks proper constraints, leading to encoder degeneration and loss of representational capacity. This fundamentally disrupts the learning of domain-invariant features. The performance drop caused by removing the CVF mechanism indicates that low-quality samples have a substantial negative impact on model performance.

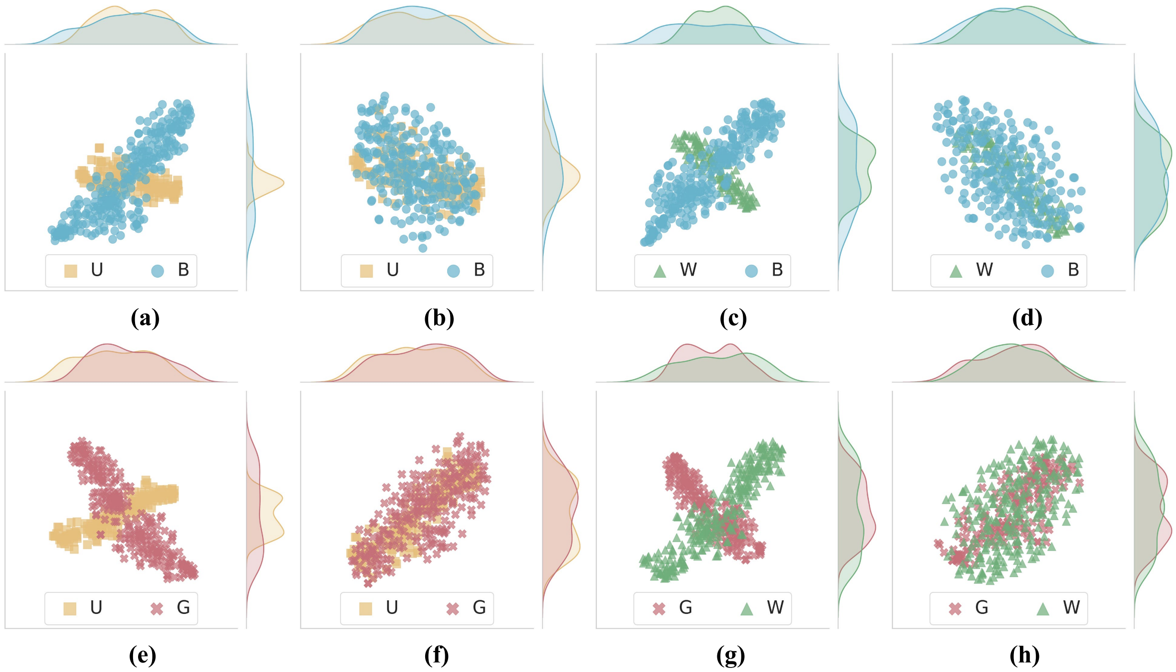

Figure 4: t-SNE visualization of the multimodal feature X and the learned domain-invariant feature F.

Sensitivity analysis

The DCC module and Test-Time Training (TTT) strategy are two key components of our proposed method. In this section, we investigate how the hyperparameter β in the $\mathcal{L}_{\text{dcc}}$ loss and the threshold θ in TTT influence the performance of VDT.

Figure 3 illustrates how VDT's performance changes as β increases from 0 to 3 with a step size of 0.5. The results show that the model achieves the best performance across different domain adaptation settings when β = 1.5. When $\beta < 1$, the model prioritizes reconstruction accuracy over disentanglement, making it difficult to learn a separable representation space. As β approaches 3, performance degrades because the model tends to ignore semantic reconstruction, which may result in representation collapse.

In addition, we evaluate the impact of the confidence threshold θ used in the CVF mechanism. We test values of θ from 0.5 to 1, and the model performance under different target domain settings is illustrated in Figure 3. The results show that the model achieves the best performance when θ = 0.9, indicating that samples with higher confidence scores tend to provide more reliable pseudo-labels. However, when the threshold exceeds 0.9, the performance drops sharply, likely due to the significant reduction in the number of selected samples.

Visualization of domain-invariant feature

To further investigate whether the DIVA module in VDT effectively aligns the source and target domains and facilitates the learning of domain-invariant features. We employ t-SNE (van der Maaten and Hinton 2008; Dimitriadis, Neto, and Kampff 2018) to project both the original multimodal features X and the learned domain-invariant features F into a 2D Euclidean space. The resulting visualizations are presented in Figure 4.

Figure 4 consists of eight subfigures grouped into four pairs: (a)-(b), (c)-(d), (e)-(f), and (g)-(h), where each pair compares the original multimodal features X and the domain-invariant features F learned through the DIVA module. A comparison within each pair reveals that the originally dispersed source and target domain features become highly overlapped in the projected space after passing through the DIVA module. This observation indicates that the distribution gap between the two domains is significantly reduced, demonstrating the effectiveness of the DIVA module in aligning cross-domain distributions and extracting domain-invariant representations.

Conclusion

In conclusion, we proposed VDT, a domain adaptive out-of-context news detection model that can aligns two domains with different distributions to adapt to the unseen target domain. The proposed approach introduces the DIVA module to jointly encode the latent distributions of source and target domains, along with a latent space alignment loss that effectively facilitates the learning of domain-invariant features. To prevent semantic collapse, a domain consistent constraint is applied to reconstruct the latent distributions. During inference, we develop CVF mechanism combined with a test-time training strategy to dynamically adapt to the target domain distribution. Extensive experiments demonstrate that our method consistently outperforms existing state-of-the-art approaches across various cross-domain settings, showing strong generalization and robustness.

References

Aneja, S.; Bregler, C.; and Nießner, M. 2021. Cosmos: Catching out-of-context misinformation with self-supervised learning. *arXiv preprint arXiv:2101.06278*.

Bhattacharjee, A.; Kumarage, T.; Moraffah, R.; and Liu, H. 2023. ConDA: Contrastive Domain Adaptation for AI-generated Text Detection. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 598–610.

Burgess, C. P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; and Lerchner, A. 2018. Understanding disentangling in $beta$-VAE. *arXiv preprint arXiv:1804.03599*.

Chen, C.; and Shu, K. 2023. Can LLM-Generated Misinformation Be Detected? In *NeurIPS 2023 Workshop on Regulatable ML*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PmLR.

Dimitriadis, G.; Neto, J. P.; and Kampff, A. R. 2018. t-SNE visualization of large-scale neural recordings. *Neural computation*, 30(7): 1750–1774.

Dolhansky, B.; Bitton, J.; Pflaum, B.; Lu, J.; Howes, R.; Wang, M.; and Ferrer, C. C. 2020. The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Farahani, A.; Voghoei, S.; Rasheed, K.; and Arabnia, H. R. 2021. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, 877–894.

Fazio, L. 2020. Out-of-context photos are a powerful low-tech form of misinformation. *The Conversation*, 14: 1.

Gandelsman, Y.; Sun, Y.; Chen, X.; and Efros, A. 2022. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems*, 35: 29374–29385.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Gu, Y.; Zhang, M.; Castro, I.; Wu, S.; and Tyson, G. 2024. Learning Domain-Invariant Features for Out-of-Context News Detection. *arXiv preprint arXiv:2406.07430*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Hershey, J. R.; and Olsen, P. A. 2007. Approximating the Kullback Leibler divergence between Gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, IV–317. IEEE.

Hu, B.; Sheng, Q.; Cao, J.; Shi, Y.; Li, Y.; Wang, D.; and Qi, P. 2024. Bad actor, good advisor: Exploring the role of large language models in fake news detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 22105–22113.

Huang, M.; Jia, S.; Chang, M.-C.; and Lyu, S. 2022. Text-image de-contextualization detection using vision-language models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8967–8971. IEEE.

Jaiswal, A.; Sabir, E.; AbdAlmageed, W.; and Natarajan, P. 2017. Multimedia semantic integrity assessment using joint embedding of images and text. In *Proceedings of the 25th ACM international conference on Multimedia*, 1465–1471.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Komodakis, N.; and Gidaris, S. 2018. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*.

Lee, D.-H.; et al. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, 896. Atlanta.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, Y.; Lee, K.; Kordzadeh, N.; Faber, B.; Fiddes, C.; Chen, E.; and Shu, K. 2021. Multi-source domain adaptation with weak supervision for early fake news detection. In *2021 IEEE International Conference on Big Data (Big Data)*, 668–676. IEEE.

Liu, F.; Wang, Y.; Wang, T.; and Ordonez, V. 2021a. Visual News: Benchmark and Challenges in News Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6761–6771.

Liu, Y.; Kothari, P.; Van Delft, B.; Bellot-Gurlet, B.; Mordan, T.; and Alahi, A. 2021b. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34: 21808–21820.

Lu, W.; Tong, Y.; and Ye, Z. 2025. DAMMFND: Domain-Aware Multimodal Multi-view Fake News Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 559–567.

Luo, G.; Darrell, T.; and Rohrbach, A. 2021. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6801–6817.

Mosallanezhad, A.; Karami, M.; Shu, K.; Mancenido, M. V.; and Liu, H. 2022. Domain adaptive fake news detection via reinforcement learning. In *Proceedings of the ACM web conference 2022*, 3632–3640.

Mu, M.; Das Bhattacharjee, S.; and Yuan, J. 2023. Self-supervised distilled learning for multi-modal misinformation identification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2819–2828.

Müller-Budack, E.; Theiner, J.; Diering, S.; Idahl, M.; and Ewerth, R. 2020. Multimodal analytics for real-world news using measures of cross-modal entity consistency. In *Proceedings of the 2020 international conference on multimedia retrieval*, 16–25.

Osowiechi, D.; Hakim, G. A. V.; Noori, M.; Cheraghalikhani, M.; Ben Ayed, I.; and Desrosiers, C. 2023. Tttflow: Unsupervised test-time training with normalizing flow. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2126–2134.

Pan, L.; Chen, W.; Kan, M.-Y.; and Wang, W. Y. 2023. Attacking Open-domain Question Answering by Injecting Misinformation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 525–539.

Papadopoulos, S.-I.; Koutlis, C.; Papadopoulos, S.; and Petrantonakis, P. C. 2023. RED-DOT: Multimodal Fact-checking via Relevant Evidence Detection. *arXiv preprint arXiv:2311.09939*.

Papadopoulos, S.-I.; Koutlis, C.; Papadopoulos, S.; and Petrantonakis, P. C. 2025. Similarity over Factuality: Are we making progress on multimodal out-of-context misinformation detection? In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 5041–5050. IEEE.

Park, J.; Kim, H.; Kang, Y.; Lim, Y.; and Kim, J. 2024. From data to discovery: recent trends of machine learning in metal–organic frameworks. *JACS Au*, 4(10): 3727–3743.

Qi, P.; Yan, Z.; Hsu, W.; and Lee, M. L. 2024. SNIFFER: Multimodal Large Language Model for Explainable Out-of-Context Misinformation Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13052–13062.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; and Hardt, M. 2020. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, 9229–9248. PMLR.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

van der Maaten, L.; and Hinton, G. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9: 2579–2605.

Wang, F.; Han, Z.; Gong, Y.; and Yin, Y. 2022. Exploring domain-invariant parameters for source free domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7151–7160.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, 849–857.

Wu, Y.; Chi, Z.; Wang, Y.; Plataniotis, K. N.; and Feng, S. 2024. Test-time domain adaptation by learning domain-aware batch normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15961–15969.

You, K.; Long, M.; Cao, Z.; Wang, J.; and Jordan, M. I. 2019. Universal domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2720–2729.

Yuan, H.; Zheng, J.; Ye, Q.; Qian, Y.; and Zhang, Y. 2021. Improving fake news detection with domain-adversarial and graph-attention neural network. *Decision Support Systems*, 151: 113633.

Yuan, X.; Guo, J.; Qiu, W.; Huang, Z.; and Li, S. 2023. Support or Refute: Analyzing the Stance of Evidence to Detect Out-of-Context Mis-and Disinformation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yue, Z.; Zeng, H.; Kou, Z.; Shang, L.; and Wang, D. 2022. Contrastive domain adaptation for early misinformation detection: A case study on covid-19. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 2423–2433.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115.

Zhang, F.; Liu, J.; Zhang, Q.; Sun, E.; Xie, J.; and Zha, Z.-J. 2023a. ECENet: Explainable and Context-Enhanced Network for Muti-modal Fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1231–1240.

Zhang, H.; Qian, S.; Fang, Q.; and Xu, C. 2020. Multimodal disentangled domain adaption for social media event rumor detection. *IEEE Transactions on Multimedia*, 23: 4441–4454.

Zhang, Y.; Liu, T.; Long, M.; and Jordan, M. 2019. Bridging theory and algorithm for domain adaptation. In *International conference on machine learning*, 7404–7413. PMLR.

Zhang, Y.; Trinh, L.; Cao, D.; Cui, Z.; and Liu, Y. 2023b. Interpretable Detection of Out-of-Context Misinformation with Neural-Symbolic-Enhanced Large Multimodal Model. *arXiv preprint arXiv:2304.07633*.