VaeDiff-DocRE: End-to-end Data Augmentation Framework for Document-level Relation Extraction

Khai Phan Tran♠, Wen Hua♣, Xue Li♠

♠ School of Electrical Engineering and Computer Science, The University of Queensland, Australia
♣ Department of Data Science and Artificial Intelligence, The Hong Kong Polytechnic University, Hong Kong SAR, China
Correspondence: phankhai.tran@uq.edu.au

Abstract

Document-level Relation Extraction (DocRE) aims to identify relationships between entity pairs within a document. However, most existing methods assume a uniform label distribution, resulting in suboptimal performance on real-world, imbalanced datasets. To tackle this challenge, we propose a novel data augmentation approach using generative models to enhance data from the embedding space. Our method leverages the Variational Autoencoder (VAE) architecture to capture all relation-wise distributions formed by entity pair representations and augment data for underrepresented relations. To better capture the multi-label nature of DocRE, we parameterize the VAE's latent space with a Diffusion Model. Additionally, we introduce a hierarchical training framework to integrate the proposed VAE-based augmentation module into DocRE systems. Experiments on two benchmark datasets demonstrate that our method outperforms state-of-the-art models, effectively addressing the long-tail distribution problem in DocRE. Our code is released at: https://github.com/khaitran22/VaeDiff -DocRE

1 Introduction

The task of relation extraction (RE) involves identifying possible relationships between two entities within a given context. This is a fundamental task in the field of Information Extraction, as it serves as a foundation for other tasks such as Question Answering (Xu et al., 2016). Early approaches to RE focused on sentence-level settings, where the context is limited to a single sentence. However, this approach is often impractical, as many relationships can only be inferred from information spread across multiple sentences. For instance, at least 40.7% of relational facts on Wikipedia can only be identified from multiple sentences (Yao et al., 2019). As a result, there has been growing
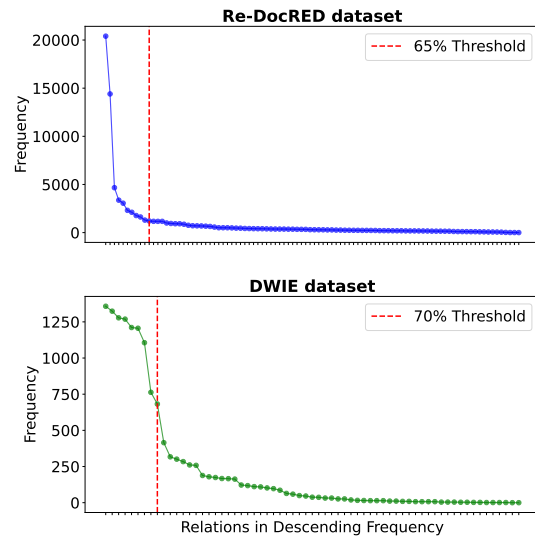

Figure 1: Relation frequency in Re-DocRED (Tan et al., 2022b) and DWIE (Zaporojets et al., 2021) datasets.

interest in exploring the problem of extracting relations across multiple sentences, a task known as Document-level Relation Extraction (DocRE) (Xu et al., 2021; Zhang et al., 2021; Zhou et al., 2021; Tan et al., 2022a).

Notably, the distribution of relations in datasets for the DocRE task often exhibits significant imbalance, where a few relation labels dominate the majority of instances. This imbalance is apparent both in the disparity between entity pairs that express at least one positive relation and those that do not, a situation we refer to as *positive-negative imbalance* (PN imbalance), and among the relation labels themselves, which we call *positive-positive imbalance* (PP imbalance). For example, in the DocRED dataset (Yao et al., 2019), at least 94% of entity pairs do not express any relation (PN imbalance). Additionally, 10 out of 96 relation labels account for more than 60% of the training instances (PP imbalance), while in the DWIE dataset (Zaporojets et al., 2021), 9 out of 66 relation labels account for more than 70% of the training instances, as illus-

trated in Figure 1. Consequently, training DocRE systems on these datasets often leads to degraded performance due to the skewed distribution.

Recent research in the DocRE task has concentrated on methods to address distribution imbalance (Zhang et al., 2021; Zhou et al., 2021; Tan et al., 2022a; Zhou and Lee, 2022; Guo et al., 2023; Wang et al., 2023). However, these studies primarily focus on mitigating the PN imbalance, often overlooking the PP imbalance phenomenon. Only a few works, notably Tan et al. (2022a) and Guo et al. (2023), have specifically investigated the PP imbalance issue. They proposed different objective functions aimed at enhancing the distinction between entity pair representations across different relation labels. However, these approaches tend to improve the performance of one type of class at the expense of another (Zhang et al., 2023). Empirically, we observe that the objective function in Tan et al. (2022a) enhances performance on minority classes but at the cost of majority classes. Additionally, removing the modified version of supervised contrastive loss proposed by Guo et al. (2023) — intended to address the PP imbalance issues — can paradoxically lead to better overall performance. A feasible solution to addressing the trade-off and PP imbalance is to increase the number of training instances for minority classes.

Inspired by this, we propose a novel data augmentation approach in the DocRE task. Since pre-processing and labeling datasets for the DocRE task are resource-intensive and prone to human errors (e.g., false negatives), we propose a data augmentation technique in the embedding space for DocRE. Specifically, we observe that after the training process, the representation of entity pairs forms distinct relation-wise clusters, as shown in Figure 2. This observation suggests that each relation has its own underlying distribution, shaped by the embeddings of entity pairs expressing that relation. By learning these distributions, we can generate entity pair representations that accurately convey the targeted relations within the embedding space.

Building on this observation, we propose a deep generative model approach to augment data for the DocRE task. This approach integrates a Variational Auto-Encoder (VAE) (Kingma and Welling, 2013) with a Diffusion Probabilistic Model (DPM) (Ho et al., 2020). Assuming that each relation exhibits a distinct distribution shaped by the entity pair representations expressing that relation, as illustrated in
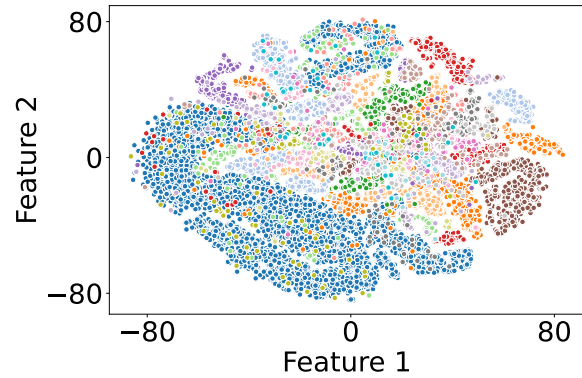

Figure 2: Relation-wise distribution from Re-DocRED dataset visualized by t-SNE (Van der Maaten and Hinton, 2008).

Figure 2, we train the VAE to capture these relation-wise distributions in the embedding space. The DPM architecture is then employed to parameterize the latent space of the VAE, effectively addressing the multi-label nature of the DocRE task. We call this module as VaeDiff. Finally, we introduce a hierarchical training framework that learns the relation-wise distributions, trains VaeDiff, and integrates it to improve the performance of the DocRE model, particularly under long-tail distribution settings.

We name our overall framework as **V**ariational **A**utoencoder and **Diff**usion Prior for **Doc**ument-level **R**elation **E**xtraction (VaeDiff-DocRE). We conduct the experiments on two DocRE dataset - the Re-DocRED (Tan et al., 2022b) and the DWIE (Zaporojets et al., 2021), which demonstrate that our VaeDiff-DocRE model significantly outperforms the state-of-the-art methods in addressing long-tail issues in DocRE task. The contribution of our work are summarized as follows:

- We propose a data augmentation framework via embedding space to address the PP imbalance and the performance trade-off issues of existing methods.

- We propose leveraging VAE and DPM architectures in our framework for augmenting data. To the best of our knowledge, we are the first applying DPM to the DocRE task.

- Extensive experiments on two public DocRE datasets, Re-DocRED (Tan et al., 2022b) and DWIE (Zaporojets et al., 2021), demonstrate the effectiveness of our VaeDiff-DocRE framework compared to state-of-the-art baselines.

2 Methodology

In this section, we introduce our VaeDiff-DocRE framework, a novel data augmentation approach at the embedding space to assist the DocRE task. Section 2.2 introduces the data augmentation module VaeDiff, which comprises a VAE architecture to model relation-wise distributions in the embedding space and a DPM to parameterize the prior distribution of the VAE, specifically designed for multi-label settings. Section 2.3 presents VaeDiff-DocRE, an end-to-end training framework that learns the relation-wise distributions, trains VaeDiff, and integrates it to enhance the performance of the DocRE model.

2.1 Problem Formulation

Let $D = \{w_l\}_{l=1}^{L}$ be a document containing L words and a set of entities $\mathcal{E}_D = \{e_i\}_{i=1}^{|\mathcal{E}_D|}$. Each entity e_i is associated with a set of mentions $\mathcal{M}_{e_i} = \{m_j^i\}_{j=1}^{|\mathcal{M}_{e_i}|}$ (i.e., a set of phrases referring to the same entity e_i). In DocRE task, we predict the subset of relations in a predefined set $\mathcal{R} = \{r_k\}_{k=1}^{|\mathcal{R}|}$ that hold between each pair of entities $(e_h, e_t)_{h,t=1,\ldots,|\mathcal{E}_D|, h \neq t}$. We sometimes abbreviate an entity pair (e_h, e_t) as (h, t) to simplify notation. A relation is deemed to exist between the head entity e_h and tail entity e_t if it is expressed between any of their corresponding mentions. If no relation exists between any pair of their mentions, the entity pair is labeled NA. For each entity pair, we term a relation that holds between its constituent entities as *positive*, and the remaining relations in $\mathcal{R}$ as *negative*. An entity pair that is NA does not have any positive relation, and has the entire set $\mathcal{R}$ as negative relations (we could consider such a pair as having a special NA relation between them).

2.2 Data Augmentation Module - VaeDiff

The proposed VaeDiff module consists of a VAE architecture, called the Entity Pair VAE (EP-VAE), designed to learn relation-wise distributions with a compact latent space. We then parameterize the latent space in EP-VAE using a diffusion model, referred to as the Diffusion Prior, specifically tailored for multi-label settings in DocRE. During inference, the Diffusion Prior generates relation-specific latent variables, which are subsequently decoded by the EP-VAE to generate new entity pair representations corresponding to the target relation.
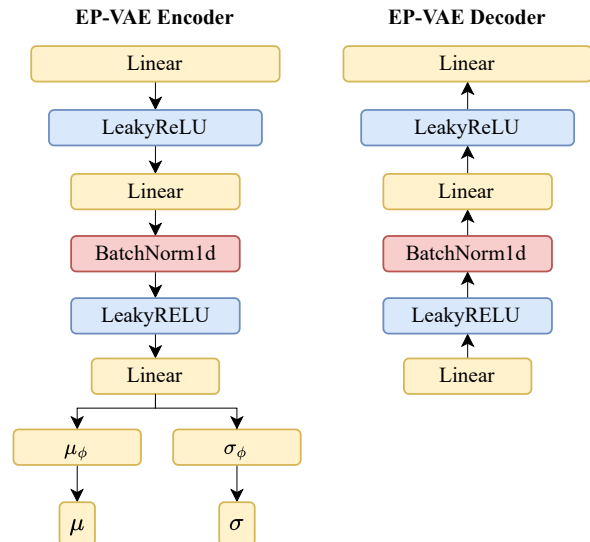

Figure 3: Structure of EP-VAE module.

2.2.1 Entity Pair VAE

The Entity Pair VAE, referred to as EP-VAE, consists of an encoder and a decoder $f_{\boldsymbol{\theta}}(\cdot)$, parameterized by $\boldsymbol{\theta}$. The encoder includes two components: (1) a projector $f_{\boldsymbol{\phi}}(\cdot)$ that maps the input to the EP-VAE latent space, and (2) separate projectors for the mean (μ_{ϕ}) and log variance (σ_{ϕ}), which estimate the parameters of the posterior distribution in the latent space. Both the encoder projector and the decoder share an identical architecture comprising alternating feedforward layers and LeakyReLU activation functions, with a single BatchNorm layer (Ioffe and Szegedy, 2015) incorporated mid-stack. The mean (μ_{ϕ}) and log variance (σ_{ϕ}) projectors are implemented as feed-forward networks. The detailed structure of EP-VAE is depicted in Figure 3.

Let $\mathcal{P} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n)$ represent the set of training entity pair representations, and let $\mathbf{z}_p$ denote the latent vectors. The projector $f_{\boldsymbol{\phi}}(\cdot)$ first maps the input $\mathbf{p}$ to a latent space representation $\mathbf{p}'$ of the EP-VAE framework. Subsequently, the mean (μ_{ϕ}) and log variance (σ_{ϕ}) projectors produce the parameters $\boldsymbol{\mu}_p$ and $\log(\boldsymbol{\sigma}_p)$, which define the posterior distribution $q_{\phi}(\mathbf{z}_p | \mathbf{p})$:

$$
\begin{aligned}
\mathbf{p}' &= f_{\boldsymbol{\phi}}(\mathbf{p}) \\
\boldsymbol{\mu}_p &= \mu_{\phi}(\mathbf{p}') \\
\log(\boldsymbol{\sigma}_p) &= \sigma_{\phi}(\mathbf{p}')
\end{aligned}
\tag{1}
$$

where $\boldsymbol{\mu}_p$ and $\boldsymbol{\sigma}_p$ are parameters of multivariate Gaussian distribution. To enable back-propagation, we use reparameterization trick proposed by
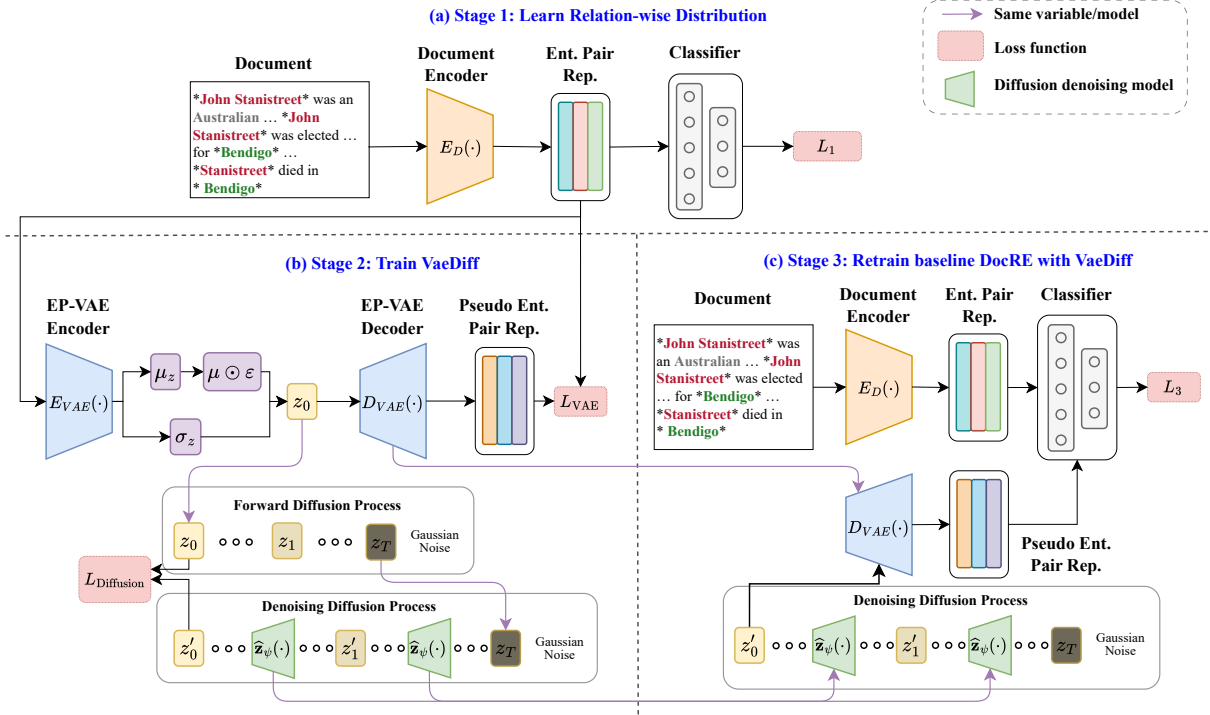
Figure 4: Overview of our VaeDiff-DocRE framework.

Kingma and Welling (2013) as follows:

$$\mathbf{z}_p = \boldsymbol{\mu}_p + \boldsymbol{\sigma}_p \odot \boldsymbol{\epsilon}, \text{ where } \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \quad (2)$$

Then, $\mathbf{z}_p$ are given to the decoder $f_{\boldsymbol{\theta}}(\cdot)$ to reconstruct the original entity pair representation as follows:

$$\tilde{\mathbf{p}} = f_{\boldsymbol{\theta}}(\mathbf{z}_p) \quad (3)$$

Finally, the EP-VAE is trained to minimize the negative Evidence Lower BOund (ELBO) (Kingma and Welling, 2013) as Eq. (4):

$$\mathcal{L}_{\text{VAE}} = -\sum_{i=1}^{n} \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{p})} \left[\log(f_{\boldsymbol{\theta}}(\mathbf{p}_i|\mathbf{z}_{\mathbf{p}_i}))\right]$$
$$- D_{\text{KL}} \left[q_{\phi}(\mathbf{z}_{\mathbf{p}_i}|\mathbf{p}_i)||p(\mathbf{z}_{\mathbf{p}_i})\right] \quad (4)$$

where $p(\mathbf{z}_{\mathbf{p}})$ is the prior distribution of the latent variable $\mathbf{z}_{\mathbf{p}}$.

The first term is the reconstruction loss, corresponding to the cross-entropy between the original entity pair representation $\mathbf{p}_i$ and its reconstruction $\tilde{\mathbf{p}}_i$. This term ensures the accurate reconstruction of each entity pair representation $\mathbf{p}_i \in \mathcal{P}$. The second term is the Kullback-Leibler divergence (D_{KL}) between the prior distribution of the latent vectors $p(\mathbf{z}_{\mathbf{p}})$ and the posterior distribution $q_{\phi}(\mathbf{z}_{\mathbf{p}}|\mathbf{p})$. This term acts as a regularization factor, encouraging the posterior distribution of the entity pair representation to remain close to the prior.

2.2.2 Diffusion Prior

In the original VAE formulation, the prior distribution $p(\mathbf{z}_{\mathbf{p}})$ is assumed to follow a standard normal distribution, which limits the latent space's ability to model the context in a multi-label scenario. This can reduce the effectiveness of the generated entity pair representations for target relations. To address this, we introduce a Diffusion Prior by parameterizing $p(\mathbf{z}_{\mathbf{p}})$ with a DPM, enhancing the complexity of the latent space in the EP-VAE. This new prior is denoted as $p_{\psi}(\mathbf{z}_{\mathbf{p}})$, where ψ represents the model's parameters. Additional background on DPM can be found in Appendix A.2.

Architecture. Our denoising network $\hat{\mathbf{z}}_{\psi}(\cdot)$ is a pre-LayerNorm transformer (Xiong et al., 2020). Each layer is conditioned on the timestep t by incorporating the Transformer sinusoidal position embedding (Vaswani et al., 2017), followed by an MLP with a single hidden layer.

Objective Function. Unlike the approach proposed by Ho et al. (2020), we discovered that training our denoising network $\hat{\mathbf{z}}_{\psi}(\mathbf{z}_t, t)$ to directly predict the original latent vectors $\tilde{\mathbf{z}}_0$ from its arbitrarily noisified version $\mathbf{z}_t$ yields better results. Thus, we train our Diffusion Prior with the objective function as below:

$$\mathcal{L}_{\text{Diff}} = \|\hat{\mathbf{z}}_{\psi}(\mathbf{z}_t, t) - \mathbf{z}_0\|_2^2 \quad (5)$$

Self-condition To improve denoising performance, we implement self-conditioning technique introduced by Chen et al. (2023). Normally, the network is conditioned on the latent variable and timestep: $\tilde{\mathbf{z}}_0^t = \hat{\mathbf{z}}_\psi(\mathbf{z}_t, t)$. Self-conditioning adds an estimate from the previous timestep s: $\tilde{\mathbf{z}}_0^t = \hat{\mathbf{z}}_\psi(\mathbf{z}_t, t, \tilde{\mathbf{z}}_0^s)$, enhancing the current prediction.

During training, the network alternates between without $\tilde{\mathbf{z}}_0^{t,\emptyset} = \hat{\mathbf{z}}_\psi(\mathbf{z}_t, t, \emptyset)$ and with self-conditioning by probability p, where $\tilde{\mathbf{z}}_0^{t,\emptyset} = \hat{\mathbf{z}}_\psi(\mathbf{z}_t, t, \emptyset)$ is computed first and refined to $\tilde{\mathbf{z}}_0^t = \hat{\mathbf{z}}_\psi(\mathbf{z}_t, t, \text{sg}(\tilde{\mathbf{z}}_0^{t,\emptyset}))$ with $\text{sg}(\cdot)$ denotes the stop-gradient operation. A learnable embedding is used when no previous estimate is available, like in the without self-conditioning.

Class Condition For multi-label settings, we propose leveraging a conditional diffusion model using the Classifier-Free Diffusion Guidance paradigm (Ho and Salimans, 2021). The model alternates between conditional generation $\tilde{\mathbf{z}}_0^t = \hat{\mathbf{z}}_\psi(\mathbf{z}_t, t, y)$ and unconditional generation $\tilde{\mathbf{z}}_0^{t,\emptyset} = \hat{\mathbf{z}}_\psi(\mathbf{z}_t, t, \emptyset)$, with probability p. The inference output $\tilde{\mathbf{z}}_0^t$ is a weighted combination of conditional and unconditional outputs with a hyper-parameter w:

$$\tilde{\mathbf{z}}_0^t = (1 - w) \cdot \tilde{\mathbf{z}}_0^{t,y} + w \cdot \tilde{\mathbf{z}}_0^{t,\emptyset} \tag{6}$$

We introduce a learnable embedding $\mathbf{y} \in \mathbb{R}^{|\mathcal{R}| \times d}$ for each label, and for multi-label cases, the conditional input is the sum of embeddings for all ground truth labels:

$$y = \sum_{i \in \mathcal{R}} \mathbb{I}(y_i = 1) \cdot \mathbf{y}_i \tag{7}$$

We also incorporate conditional information into the time embedding.

2.2.3 Training Objective

To co-train Diffusion Prior and EP-VAE, we decompose the Kullback-Leibler divergence term D_{KL} in Eq. (4) into its negative encoder entropy and cross-entropy. Specifically, the decomposition is as follows:

$$D_{\text{KL}}\left[q_\phi(\mathbf{z}_{\mathbf{p}}|\mathbf{p})||p_\psi(\mathbf{z}_{\mathbf{p}})\right] = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{p}}|\mathbf{p})}\left[\log(q_\phi(\mathbf{z}_{\mathbf{p}}|\mathbf{p}))\right]}_{\text{negative encoder entropy}}$$
$$+ \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{p}}|\mathbf{p})}\left[-\log(p_\psi(\mathbf{z}_{\mathbf{p}}))\right]}_{\text{cross-entropy}} \tag{8}$$

Following Wehenkel and Louppe (2021), the cross-entropy term can further be replaced by the regression objective $\mathcal{L}_{\text{Diff}}$ of the diffusion model in Eq.

(5). As a result, the VaeDiff component is trained with the following objective function:

$$\mathcal{L}_{\text{VaeDiff}} = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{p}}|\mathbf{p})}\left[-\log(f_{\boldsymbol{\theta}}(\mathbf{p}|\mathbf{z}_{\mathbf{p}}))\right]}_{\text{reconstruction EP-VAE loss}}$$
$$+ \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{p}}|\mathbf{p})}\left[\log(q_\phi(\mathbf{z}_{\mathbf{p}}|\mathbf{p}))\right]}_{\text{negative EP-VAE encoder entropy}} \tag{9}$$
$$+ \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_{\mathbf{p}}|\mathbf{p})}\left[\|\hat{\mathbf{z}}_\psi(\mathbf{z}_t, t) - \mathbf{z}_0\|_2^2\right]}_{\mathcal{L}_{\text{Diff}}}$$

The detailed derivation of Eq. (9) is presented in Appendix A.4.

2.3 Hierarchical Training Data Augmentation Framework - VaeDiff-DocRE

In this section, we introduce VaeDiff-DocRE, an end-to-end hierarchical three-stage training data augmentation framework designed to improve the performance of the DocRE system under long-tail distribution settings. The overall framework is depicted in Figure 4.

2.3.1 Stage 1: Learn Relation-wise Distribution

To obtain the structured embedding space depicted in Figure 2, we train a simple baseline DocRE model, which initially performs lower than the compared baseline methods. The primary goal of this baseline is to mitigate the negative impact of the PN imbalance, which is not the main focus of this work, and to demonstrate the effectiveness of our proposed data augmentation approach. This baseline DocRE model can be replaced by any more advanced DocRE model to achieve better results. For entity pair representation learning, we adopt the structure of KD-DocRE (Tan et al., 2022a) as the backbone for our baseline model. Regarding the objective function, we replace the Adaptive Focal Loss used in KD-DocRE with PMTEM loss (Guo et al., 2023) and Supervised Contrastive Learning loss (Khosla et al., 2020). The detailed structure of this baseline model is provided in Appendix A.3.

2.3.2 Stage 2: Train VaeDiff Module

Following the training of the baseline DocRE model in stage 1, we proceed to train the VaeDiff augmentation module as described in Section 2.2. The training inputs for the VaeDiff module are generated by utilizing the document encoder from the trained DocRE baseline to extract entity pair representations within a document. Initially, the EP-

VAE is trained independently, without co-training the Diffusion Prior, to allow the latent space to stabilize over the first n epochs. Subsequently, the EP-VAE and the Diffusion Prior are co-trained using the objective function specified in Eq. (9).

2.3.3 Stage 3: Retrain baseline DocRE with VaeDiff

In the final stage, we retrain the baseline DocRE model using the trained VaeDiff augmentation model to address the PP imbalance in the DocRE task. The detailed algorithm of stage 3 is provided in Algorithm 1.

Following the warmup approach in stage 2, we initially train the baseline DocRE model for the first n epochs without augmenting any new entity representations. After training for initial n epochs, we then train the DocRE model using both the features $\mathbf{x}$ from its document encoder and the pseudo-features $\tilde{\mathbf{x}}$ generated by the VaeDiff model.

To generate $\tilde{\mathbf{x}}$, assume there are $|\mathcal{E}_p|$ positive entity pairs in the current batch $\mathcal{B}$. We select the label vectors $l_{\mathcal{E}_p}$ of these positive entity pairs. The VaeDiff model then generates $m \times |\mathcal{E}_p|$ additional entity pair representations using $l_{\mathcal{E}_p}$ as class conditional information. The DocRE model makes predictions based on the combined set $\mathbf{X} = \{\mathbf{x}, \tilde{\mathbf{x}}\}$, and these predictions are concatenated to compute the loss for training the DocRE model.

3 Experiments

3.1 Experimental Setup

Dataset. We evaluated our proposed method using two benchmark datasets: Re-DocRED (Tan et al., 2022b) and DWIE (Zaporojets et al., 2021). Re-DocRED is a high-quality, revised version of DocRED (Yao et al., 2019) – a large-scale document-level RE dataset constructed from Wikipedia and Wikidata, which supplements previously missed false negative triplets. DWIE is a document-level information extraction dataset that includes relation extraction. The statistics for these two datasets are provided in Table 3.

Metrics. Following Yao et al. (2019), we use micro F1 (**F1**) and micro Ignore F1 (**Ign F1**) as evaluation metrics to assess overall DocRE performance. Ign F1 measures the F1 score while excluding relational triplets shared between the training and test sets. Additionally, we use two F1 scores specific to long-tailed distribution settings proposed by Tan et al. (2022b): **Freq. F1** for common relations

---

Algorithm 1 `VaeDiff-DocRE` framework

Input: Dataset $\mathcal{D}$
Model: DocRE model $f_\xi(\cdot)$, Diffusion Prior model $\hat{\mathbf{z}}_\psi(\cdot)$, EP-VAE Decoder $f_\theta(\cdot)$

1: **while** epoch $\leq$ total_epochs **do**
2:     $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}$ ▷ Sample batch data
3:     **if** epoch $\leq$ warmup_epochs **then**
4:         $\hat{\mathbf{y}} \leftarrow f_\xi(\mathbf{x})$ ▷ Predict relation labels
5:         Compute $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$ by Eq (29).
6:
7:     **else**   ▷ Ent. Pair Rep Augmentation Process
8:         $\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ ▷ Sample noise
9:         **for** $t = T, ..., 1$ **do**   ▷ Denoising process
10:
11:             // Self-conditioning
12:             **if** $t == T$ **then**
13:                 $\tilde{\mathbf{z}}_0^{t,y} \leftarrow \hat{\mathbf{z}}_\psi(\mathbf{z}_t, t, \emptyset, \mathbf{y})$
14:                 $\tilde{\mathbf{z}}_0^{t,\emptyset} \leftarrow \hat{\mathbf{z}}_\psi(\mathbf{z}_t, t, \emptyset, \emptyset)$
15:             **else**
16:                 $\tilde{\mathbf{z}}_0^{t,y} \leftarrow \hat{\mathbf{z}}_\psi(\mathbf{z}_t, t, \text{sg}(\tilde{\mathbf{z}}_0^{t+1}), \mathbf{y})$
17:                 $\tilde{\mathbf{z}}_0^{t,\emptyset} \leftarrow \hat{\mathbf{z}}_\psi(\mathbf{z}_t, t, \text{sg}(\tilde{\mathbf{z}}_0^{t+1}), \emptyset)$
18:             **end if**
19:
20:             // Class-conditioning
21:             $\tilde{\mathbf{z}}_0^t = (1 - w) \cdot \tilde{\mathbf{z}}_0^{t,y} + w \cdot \tilde{\mathbf{z}}_0^{t,\emptyset}$
22:
23:             // Sample sampling
24:             $\tilde{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$ else $\tilde{\epsilon} = 0$
25:             Compute $\tilde{\boldsymbol{\mu}}_t$ as in DDPM sampling
26:             $\mathbf{z}_{t-1} = \tilde{\boldsymbol{\mu}}_t + \sigma_t * \tilde{\epsilon}$
27:
28:         **end for**
29:         $\tilde{\mathbf{x}} \leftarrow f_\theta(\tilde{\mathbf{z}}_0)$ ▷ Generating New Rep
30:
31:         // Prediction
32:         $\hat{\mathbf{y}}, \tilde{\mathbf{y}} \leftarrow f_\xi(\mathbf{x}), f_\xi(\tilde{\mathbf{x}})$
33:         $\hat{\mathbf{y}}' = \text{Concat}([\tilde{\mathbf{y}}, \hat{\mathbf{y}}])$
34:         Compute $\mathcal{L}(\hat{\mathbf{y}}', \mathbf{y})$ by Eq (29).
35:     **end if**
36:     Update DocRE model $f_\xi(\cdot)$
37: **end while**

---

and **LTail F1** for uncommon relations. For the Re-DocRED datasets, Freq F1 is calculated for the 10 most common relations, while LT F1 covers the remaining 86 relations.

Implementation details. We implemented our method using Huggingface's Transformers (Wolf et al., 2020) and PyTorch 2.1.2 (Paszke et al., 2019). For the document encoders, we utilized BERT-base-cased (Devlin et al., 2019) and RoBERTa-large (Liu et al., 2019). The model was optimized using the AdamW optimizer (Loshchilov and Hutter, 2019). All models were implemented and trained on an NVIDIA H100 GPU. We utilized the public repositories of baseline models to implement our experiments.

In stage 1, we configure the learning rates for the document encoder at $3e - 5$ and the classifier head at $1e - 4$ for the DocRED and Re-DocRED

Table 1 columns: Method | PLM | Dev: Ign F1, F1, Freq F1, LTail F1 | Test: Ign F1, F1, Freq F1, LTail F1

Method	PLM	Ign F1	F1	Freq F1	LTail F1	Ign F1	F1	Freq F1	LTail F1
LLM-based									
GPT-3.5 + NLI (Li et al., 2023)	GPT-3.5	-	-	-	-	-	9.33	-	-
ZeroDocRTE (Sun et al., 2024)	LLaMA2-13B-Chat	-	11.90	-	-	-	13.23	-	-
ZeroDocRE (Sun et al., 2024)	LLaMA2-13B-Chat	-	48.51	-	-	-	49.21	-	-
AutoRE (Xue et al., 2024)	Vicuna-7B	-	54.29	-	-	-	53.84	-	-
LMRC (Li et al., 2024)	LLaMA2-13B-Chat	-	-	-	-	74.08	74.63	-	-
CNN (Yao et al., 2019)[†]	-	53.95	55.60	-	-	52.80	54.88	-	-
LSTM (Yao et al., 2019)[†]	-	56.40	58.30	-	-	56.31	57.83	-	-
BiLSTM (Yao et al., 2019)[†]	-	58.20	60.04	-	-	57.84	59.93	-	-
GAIN (Zeng et al., 2020)[†]	$BERT_{base}$	71.99	73.49	-	-	71.88	73.44	-	-
ATLOP (Zhang et al., 2021)	$BERT_{base}$	73.35[†]	74.22[†]	-	-	71.86[*]	72.61[*]	75.92[*]	67.46[*]
PEMSCL (Guo et al., 2023)	$BERT_{base}$	73.82	74.50	78.21	69.03	73.05	74.14	77.77	68.38
KMGRE (Jiang et al., 2022)[†]	$BERT_{base}$	73.33	74.44	-	-	73.39	74.46	-	-
NRCL (Zhou and Lee, 2022)	$BERT_{base}$	73.74	74.84	77.95	70.12	73.40	74.48	77.53	69.59
CAST (Tan et al., 2023)[*]	$BERT_{base}$	-	-	-	-	73.32	74.67	**78.53**	69.34
KD-DocRE (Tan et al., 2022a)	$BERT_{base}$	74.37	75.31	78.71	70.26	73.98	74.88	78.28	69.52
VaeDiff-DocRE (Ours)	$BERT_{base}$	**74.96±0.21**	**75.89±0.2**	**79.18±0.22**	**71.12±0.38**	**74.13±0.21**	**75.07±0.2**	78.32±0.15	**70.06±0.29**
JEREX (Eberts and Ulges, 2021)	$RoBERTa_{Large}$	71.59	72.68	-	-	71.45	72.57	77.09	66.31
ATLOP (Zhou and Lee, 2022)	$RoBERTa_{Large}$	76.79	77.46	-	-	76.82	77.56	80.78	72.29
DocuNET (Zhang et al., 2021)	$RoBERTa_{Large}$	77.49	78.14	-	-	77.26	77.87	81.25	73.32
KD-DocRE (Tan et al., 2022a)	$RoBERTa_{Large}$	77.85	78.51	-	-	77.60	78.28	80.85	74.31
VaeDiff-DocRE (Ours)	$RoBERTa_{Large}$	**78.35±0.22**	**79.19±0.16**	**82.0±0.18**	**75.13±0.35**	**78.22±0.27**	**79.03±0.21**	**81.84±0.21**	**74.73±0.32**

Table 1: Experimental result on the development and test set of Re-DocRED. **Bold** is best performance. [†], [*] are reported from Jiang et al. (2022) and Tan et al. (2023) respectively. The result on $RoBERTa_{Large}$ PLM is reported from Tan et al. (2022b). Our results are average of ten runs and conducted statistical significant test following Dror et al. (2018).

Method	Dev Ign F1	Dev F1	Test Ign F1	Test F1
CNN	37.65	47.73	34.65	46.14
LSTM	40.86	51.77	40.81	52.60
BiLSTM	40.46	51.92	42.03	54.47
Context-Aware	42.06	53.05	45.37	56.58
CorefBERT	57.18	61.42	61.71	66.59
GAIN	58.63	62.55	62.37	67.57
SSAN	58.62	64.49	62.58	69.39
ATLOP	59.03	64.82	62.09	69.94
RSMAN	60.02	65.88	63.42	70.95
Correl	61.10	65.73	65.64	71.56
VaeDiff-DocRE (ours)	**61.51±0.58**	**67.77±0.46**	**66.44±0.48**	**73.07±0.37**

Table 2: Comparison result on the DWIE dataset. The PLM encoder is BERT-base-cased (Devlin et al., 2019). Our results are average of ten runs. The results are reported from Han et al. (2024). **Bold** is best performance.

datasets, while setting $3e − 4$ for the DWIE dataset. The training process is conducted over 30 epochs. In stage 2, we set the latent size of the EP-VAE to be 384 and employ a Cosine Annealing schedule (Loshchilov and Hutter, 2022) for the learning rate, ranging from $2e − 5$ to $2e − 4$. For the Diffusion Prior, the diffusion step is set to $T = 50$, utilizing a Cosine Noise Scheduler (Nichol and Dhariwal, 2021). Our denoising architecture consists of 6 layers with a dimension of 256. During training, the probability p of self-condition and class condition is 0.1, and the hyper-parameter w in Eq. (6) is set to 0.1. The VaeDiff model is trained over 100 epochs. In stage 3, the initial training epochs n is 5 epochs on the Re-DocRED datasets and 10 epochs on the DWIE dataset. We set $m = 2$. The overall training epochs for the DocRE models are 30 for Re-DocRED and 50 for DWIE.

3.2 Baseline

Re-DocRED For the Re-DocRED dataset, we compared our model with the following baselines in addressing PP imbalance: (1) **KD-DocRE** (Tan et al., 2022a) introduces a loss function based on Focal Loss (Lin et al., 2017), specifically tailored for the DocRE task under long-tail distribution settings. (2) **PEMSCL** (Guo et al., 2023) proposes a modified supervised contrastive loss function to address similar challenges in the task. (3) **CAST** (Tan et al., 2023) combines self-training with a pseudo-label re-sampling method to improve performance on infrequent relations.

DWIE For the DWIE dataset, we compared our method with the state-of-the-art (SOTA) method in address the PP imbalance – **Correl** (Han et al., 2024). It introduces a module designed to learn the relational correlation by leveraging the semantic distance between common and uncommon relations to enhance performance on the latter.

Datasets		# Doc.	# Rel.	# Ent.	# Facts
	Train	3,053		59,359	85,932
Re-DocRED	Dev	500	96	9,684	17,284
	Test	500		9,779	17,448
	Train	602		16,494	14,403
DWIE	Dev	98	65	2,785	2,624
	Test	99		2,623	2,459

Table 3: Statistics of DWIE and Re-DocRED.

3.3 Main Results

We present the performance of our proposed VaeDiff-DocRE method for the DocRE task, benchmarking it against baseline methods in Table 1 and Table 2. To validate the significance of our improvements, we performed significance tests following the methodology outlined by Dror et al. (2018), with all p-values being less than 0.05. These tables illustrate that our VaeDiff-DocRE framework surpasses state-of-the-art methods, particularly in both overall and long-tail performance measured by the F1 score.

In the Re-DocRED dataset, as presented in Table 1, the VaeDiff-DocRE approach demonstrated superior performance compared to baseline methods, achieving an average F1 score improvement of 0.58 and 0.68 on the development set and 0.19 and 0.75 on the test set, when utilizing BERT and RoBERTa as document encoders, respectively. Furthermore, all PLM-based approaches, including ours, outperformed LLM-based methods by a substantial margin. Regarding the long-tailed distribution, particularly for minority relations, our method outperformed the SOTA method by 0.54 and 0.42 points in F1 score on the test sets. While VaeDiff-DocRE achieved the second-highest performance for majority relations on the test set when using BERT, it achieved the best performance for majority relations with RoBERTa. Consequently, VaeDiff-DocRE secured the highest overall F1 score for both PLMs. Similarly, in the DWIE dataset, as shown in Table 2, VaeDiff-DocRE surpassed the SOTA method by 1.6 and 2.27 points in F1 scores on the development and test sets, respectively. These results underscore the effectiveness of our approach in addressing the challenges posed by long-tail distributions while maintaining strong performance for majority classes, thereby enhancing overall performance in the DocRE task.

3.4 Ablation Study

We conducted an ablation study on the Re-DocRED dataset to demonstrate the effectiveness of each proposed component, and the results are presented in Table 4.

First, we removed the VaeDiff module, and instead added Gaussian noise to the representations of positive entity pairs in the current batch $\mathcal{B}$, referred to as *Gaus Noise*. The results show that simply adding Gaussian noise to the entity pair representations yields better performance compared to the KD-DocRE method. However, this basic augmentation technique performs worse than our method, further highlighting the effectiveness of the proposed approach.

Second, we did not augment new entity pair representations during training, referred to as *No Aug*. This model corresponds to the baseline DocRE model used in stage 1 (§2.3.1). As shown in the results, this led to a significant drop in performance, with an overall F1 score decrease of 0.69. However, when augmentation was introduced, performance improved for both majority and minority relations. Notably, augmenting new data increased the F1 score for minority relations by nearly 1% in the LTail F1 score, demonstrating the positive impact of data augmentation in addressing the class-imbalance issue of the DocRE task.

Model	Ign F1	F1	Freq F1	LTail F1
VaeDiff-DocRE (Ours)	**74.96**	**75.89**	**79.18**	**71.12**
- *Gaus Noise*	<u>74.51</u>	<u>75.44</u>	<u>78.80</u>	<u>70.61</u>
- *No Aug*	74.23	75.20	78.74	69.97
KD-DocRE	74.37	75.31	78.71	70.26

Table 4: Ablation study on the Re-DocRED development set. The PLM is BERT-base-cased (Devlin et al., 2019). Second highest performance is <u>underline</u>.

3.5 Analysis of Generated Features

In Figure 5, we visualize the generated representations of entity pairs expressing minority relations using our VaeDiff approach. By comparing these with the encoded features from the document encoder of the trained DocRE model on the development set of the Re-DocRED dataset, we observe a strong alignment between the generated pair representations and the encoded features. This alignment addresses the lack of high-quality data in minority classes by enriching the true entity pair representations for these relations, effectively enhancing the performance of DocRE models in long-

tail distribution settings. However, Figure 5 also shows that some entity pair representations expressing different relations overlap significantly. This overlap occurs because these relations frequently co-occur in the dataset. For instance, entity pairs expressing the "P580" relation are highly likely to also express the "P582" and "P585" relations in the Re-DocRED dataset.
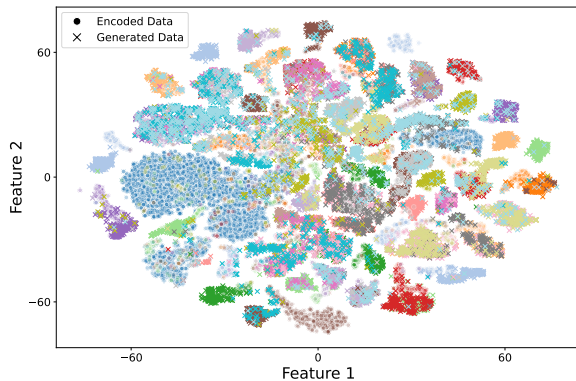

Figure 5: t-SNE visualization of relation-wise distributions of encoded and generated Entity Pair Representations by the trained VaeDiff module (§2.2). Each distribution is depicted in a unique color. ● represents the encoded actual entity pair representations within the document and × denotes the representations generated by the module.

4 Related Work

Document-level Relation Extracion (DocRE) Early works in Relation Extraction (RE) focused on sentence-level tasks, where context was confined to a single sentence. However, many relational facts require synthesizing information across multiple sentences. Since Yao et al. (2019) introduced DocRED, a large-scale document-level RE dataset, various methods have been developed for the DocRE task. One approach leverages Graph Neural Networks (GNNs) to transform documents into graph structures, as seen in the work of Nan et al. (2020) and Zeng et al. (2020). Another approach uses Transformer architectures, with Zhou et al. (2021) and Tan et al. (2022a) improving entity pair representation and reasoning through attention mechanisms.

Long-tail Issue in DocRE Real-world DocRE datasets often have a skewed distribution, where a few relations dominate most instances. This creates two challenges: an imbalance between negative and positive instances, and an imbalance among the positive instances themselves. The first issue

has been widely studied. For instance, Zhou et al. (2021) introduced an adaptive thresholding loss that treats NA logits as a learnable threshold, while Zhou and Lee (2022) reframed the task as label-ranking, and Guo et al. (2023) used pairwise ranking to better separate NA from positive relations. However, the second issue—imbalance among positive instances—has received less attention. Tan et al. (2022a) tackled it by enhancing the loss function from Zhou et al. (2021) to prioritize uncommon relations using Focal Loss (Lin et al., 2017). Similarly, Guo et al. (2023) adapted Supervised Contrastive Loss (Khosla et al., 2020) for the same goal. Yet, while these methods boost performance on rare relations, they often reduce accuracy on common classes, leading to suboptimal overall results.

Variational Autoencoder in RE To mitigate the heavy reliance on labeled data in the RE task, several studies have explored unsupervised approaches. To learn feature-rich representations that are beneficial in scenarios with limited labeled data, Marcheggiani and Titov (2016) introduced a method leveraging VAEs. In this approach, the VAE encoder is trained as a relation classifier, while the decoder reconstructs the entity pairs from the input sentences based on the encoder's classifications. This setup allows the decoder to provide a supervisory signal to the encoder. Building on this, Yuan and Eldardiry (2021) proposed a method where the relation classification r is treated as an intermediate variable conditioned on the input sentence x. In their VAE framework, the latent variable z is sampled based on the joint contributions of r and x, and the decoder uses r and z to reconstruct the entity pairs within the sentence x.

5 Conclusion

In this paper, we present VaeDiff, a novel data augmentation framework aimed at addressing long-tail distribution challenges in the DocRE task. Our approach utilizes a Variational Autoencoder (VAE) to learn relation-wise distributions and a Diffusion Probabilistic Model to enhance the EP-VAE's latent space for multi-label scenarios. We also propose a hierarchical framework VaeDiff-DocRE to boost DocRE system performance under long-tail conditions by integrating the VaeDiff module. Extensive experiments and ablation studies show that our framework outperforms SOTA methods.

Limitation

The proposed VaeDiff-DocRE framework requires multiple training rounds and inherits the same limitations as the original Diffusion Probabilistic Model framework. As a result, the GPU training time for this framework is longer compared to other models. Additionally, while the framework has been tested in the general domain, its performance in more specialized domains, such as the medical field, has yet to be evaluated.

Acknowledgements

References

Ting Chen, Ruixiang ZHANG, and Geoffrey Hinton. 2023. Analog bits: Generating discrete data using diffusion models with self-conditioning. In *The Eleventh International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1383–1392.

Markus Eberts and Adrian Ulges. 2021. An end-to-end model for entity-level relation extraction using multi-instance learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3650–3660.

Jia Guo, Stanley Kok, and Lidong Bing. 2023. Towards integration of discriminability and robustness for document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2606–2617.

Ridong Han, Tao Peng, Benyou Wang, Lu Liu, Prayag Tiwari, and Xiang Wan. 2024. Document-level relation extraction with relation correlations. *Neural Networks*, 171:14–24.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. 2019. Axial attention in multidimensional transformers. *arXiv preprint arXiv:1912.12180*.

Jonathan Ho and Tim Salimans. 2021. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 448–456. PMLR.

Feng Jiang, Jianwei Niu, Shasha Mo, and Shengda Fan. 2022. Key mention pairs guided document-level relation extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1904–1914.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Diederik Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Junpeng Li, Zixia Jia, and Zilong Zheng. 2023. Semi-automatic data enhancement for document-level relation extraction with distant supervision from large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5495–5505.

Xingzuo Li, Kehai Chen, Yunfei Long, and Min Zhang. 2024. Llm with relation classifier for document-level relation extraction. *arXiv preprint arXiv:2408.13889*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ilya Loshchilov and Frank Hutter. 2022. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*.

Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Transactions of the Association for Computational Linguistics*, 4:231–244.

Guoshun Nan, Zhijiang Guo, Ivan Sekulić, and Wei Lu. 2020. Reasoning with latent structure refinement for document-level relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1546–1557.

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qi Sun, Kun Huang, Xiaocui Yang, Rong Tong, Kun Zhang, and Soujanya Poria. 2024. Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction. In *Proceedings of the ACM on Web Conference 2024*, pages 4407–4416.

Qingyu Tan, Ruidan He, Lidong Bing, and Hwee Tou Ng. 2022a. Document-level relation extraction with adaptive focal loss and knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1672–1681.

Qingyu Tan, Lu Xu, Lidong Bing, and Hwee Tou Ng. 2023. Class-adaptive self-training for relation extraction with incompletely annotated training data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8630–8643.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022b. Revisiting docred-addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Jize Wang, Xinyi Le, Xiaodi Peng, and Cailian Chen. 2023. Adaptive hinge balance loss for document-level relation extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3872–3878.

Antoine Wehenkel and Gilles Louppe. 2021. Diffusion priors in variational autoencoders. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. 2020. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on freebase via relation extraction and textual evidence. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2326–2336.

Wang Xu, Kehai Chen, and Tiejun Zhao. 2021. Document-level relation extraction with reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14167–14175.

Lilong Xue, Dan Zhang, Yuxiao Dong, and Jie Tang. 2024. Autore: document-level relation extraction with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 211–220.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Chenhan Yuan and Hoda Eldardiry. 2021. Unsupervised relation extraction: A variational autoencoder approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1929–1938.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. Dwie: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563.

Shuang Zeng, Runxin Xu, Baobao Chang, and Lei Li. 2020. Double graph based reasoning for document-level relation extraction. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1630–1640.

Ningyu Zhang, Xiang Chen, Xin Xie, Shumin Deng, Chuanqi Tan, Mosha Chen, Fei Huang, Luo Si, and Huajun Chen. 2021. Document-level relation extraction as semantic segmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. 2023. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Conference on Empirical Methods in Natural Language Processing*.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14612–14620.

Yang Zhou and Wee Sun Lee. 2022. None class ranking loss for document-level relation extraction. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4538–4544. International Joint Conferences on Artificial Intelligence Organization.

A Background

A.1 Variational Autoencoder (VAE)

In variational autoencoder (VAE), a set of input data is denoted as $X = (\mathbf{x}_1, ..., \mathbf{x}_n)$ where n denotes the number of total input samples. The latent variables are denoted by vector $\mathbf{z}$. The VAE includes a encoder network including network and variational parameters ϕ that produces variational probability model $q_\phi(\mathbf{z}|\mathbf{x})$ and a decoder network parameterized by $\boldsymbol{\theta}$ to reconstruct sample $\tilde{\mathbf{x}} = f_{\boldsymbol{\theta}}(\mathbf{z})$. The objective of VAE is to maximize the Evidence Lower BOund (ELBO) (Kingma and Welling, 2013) as follows:

$$\max_{\boldsymbol{\theta}, \phi} \left\{ \sum_{i=1}^{n} \text{ELBO}(\boldsymbol{\theta}, \phi; \mathbf{x}_i) \right\}. \quad (10)$$

where:

$$\begin{aligned} \text{ELBO}(\boldsymbol{\theta}, \phi; \mathbf{x}) := \; & \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log(p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}))] \\ & - \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \end{aligned} \quad (11)$$

The first term in Eq. (11) corresponds to the negative reconstruction loss between input $\mathbf{x}$ and reconstruction $\tilde{\mathbf{x}}$ under Gaussian parameterization of the output. The second term in Eq. (11) refers to the KL divergence between the variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior distribution $p(\mathbf{z})$. In general, the prior distribution is assumed to be an isotrophic Gaussian distribution.

A.2 Diffusion Probabilistic Model

Given an observation of data $\mathbf{x}_0$, diffusion models (Ho et al., 2020) learn the data distribution $p(\mathbf{x}_0)$ by reversing a diffusion process. The diffusion (forward) process is a Markov chain that corrupts the sampled data $\mathbf{x}_0$ by gradually adding random noise to it:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad (12)$$

where $\beta_{1:T}$ are the pre-defined noise variances, $\beta_t \in (0, 1)$ at time step t. When $\beta_t \to T$, the data distribution will be corrupted to $\mathcal{N}(0, \mathbf{I})$. By defining $\alpha_t = \prod_{i=1}^{t}(1 - \beta_i)$, we can directly get $\mathbf{x}_t$ by adding noise to the input as follows:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0, (1 - \alpha_t)\mathbf{I}) \quad (13)$$

where $\alpha_t \in (0, 1)$.

Given access to the original data $\mathbf{x}_0$, the backward process can be inverted analytically:

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) = \mathcal{N}(f_t(\mathbf{x}_t, \mathbf{x}_0), \sigma_t^2\mathbf{I}) \quad (14)$$

where σ_t can be derived from β_t, $f_t(\mathbf{x}_t, \mathbf{x}_0)$ has a closed form parameterized by t (Ho et al., 2020). However, since the original data $\mathbf{x}_0$ is not available in the actual generation process, (i.e., the response is supposed to be generated), we can not directly use Eq. (14) to sample data and thus approximate $f_t(\cdot)$ using a neural network with the parameter ψ, namely *denoising network*.

The training of a denoising diffusion network is defined by minimizing the Kullback-Leibler divergence $D_{\text{KL}}\left[q(\mathbf{x}_{0:T}) \, || \, p_\psi(\mathbf{x}_{0:T})\right]$ between the forward trajectory $q(\mathbf{x}_{0:T})$ and the backward trajectory $p_\psi(\mathbf{x}_{0:T})$. This divergence term can be decomposed as follows:

$$\begin{aligned} D_{\text{KL}}\left[q(\mathbf{x}_{0:T})||p_\psi(\mathbf{x}_{0:T})\right] = \; & \mathbb{E}_{\mathbf{x}_0 \sim q}\left[\log q(\mathbf{x}_0)\right] - \\ & \mathbb{E}_{\mathbf{x}_{0:T} \sim q}\left[\log \frac{p_\psi(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] \end{aligned} \quad (15)$$

The second term in Eq. (15) represents the expectation of an evidence lower bound (ELBO) with respect to $\mathbf{x}_0 \sim q$:

$$\text{ELBO} := \mathbb{E}_{\mathbf{x}_{0:T} \sim q}\left[\log \frac{p_\psi(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\right] \leq \log p_\psi(\mathbf{x}_0) \quad (16)$$

By leveraging the Gaussian transitions and simplifying further, Ho et al. (2020) express the loss function in a more tractable form:

$$\mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t} \left[\frac{1}{2\sigma_t^2} \| f_\psi(\mathbf{x}_t, t) - \mathbf{x}_0 \|_2^2 \right] \qquad (17)$$

where $t \sim \text{Uniform}(\{1, \cdots, T\})$, $\mathbf{x}_t \sim q(\mathbf{x}_t|\mathbf{x}_0)$.

For inference, we can use the trained denoising network $f_\psi(\mathbf{x}_t, t)$ to build the usable inversion $p_\psi(\mathbf{x}_{t-1}|\mathbf{x}_t) \approx p(\mathbf{x}_{t-1}|\mathbf{x}_t, f_\psi(\mathbf{x}_t, t))$ and get new high-quality data by sampling from it iteratively.

A.3 Baseline DocRE

This section introduces the baseline DocRE model used in our framework introduced in Section 2.3.1.

A.3.1 Representation Learning

Encoding We insert a special token "*" before and behind each mention m_j^i of e_i in document D as an entity marker (Zhang et al., 2017). We use a pre-trained language model (PLM) to obtain token embeddings $H \in \mathbb{R}^{L \times d}$ and token-level cross-attention matrix $A \in \mathbb{R}^{L \times L}$ of D:

$$H, A = PLM([w_1, w_2, ..., w_L]) \qquad (18)$$

The representations of an entity e_i and an entity pair (e_h, e_t) are computed as Eq. (19) and Eq. (20) respectively.

$$\mathbf{e}_i = \log \sum_{i=1}^{|\mathcal{M}_{e_i}|} \exp\left(\mathbf{m}_j^i\right) \qquad (19)$$

$$\mathbf{p}_{h,t} = \tanh\left(\mathbf{W}_p \left[\mathbf{e}_h; \mathbf{e}_t; \mathbf{c}_{h,t}\right] + \mathbf{b}_p\right) \qquad (20)$$

where $\mathbf{m}_j^i$ is embedding of starting token "*" of each mention m_j^i of e_i; $\mathbf{W}_p$, $\mathbf{b}_p$ are trainable parameters; $[\ ;\ ]$ is the concatenation; $\mathbf{c}_{h,t}$ is the localized context embedding of (e_h, e_t) (Zhou et al., 2021).

Axial Attention After obtaining the representation of entity pairs, we incorporate Axial Attention (Ho et al., 2019) proposed in (Tan et al., 2022a), which allows us to integrate neighboring axial information, thereby enhancing the performance of DocRE systems.

Given a set of entities $\mathcal{E}_D = \{e_i\}_{i=1}^n$ in a document D, we construct an $n \times n$ entity table, where each row and column corresponds to an entity in D. For any entity pair (e_h, e_t), the axial attention mechanism attends to the pair's axial elements, which can be either (e_h, e_i) or (e_i, e_t). Since multi-hop reasoning is crucial in the DocRE task, if

(e_h, e_t) expresses a two-hop relation, axial attention enables the model to explore intermediate relations by connecting through a bridge entity e_i. Specifically, axial attention is computed by applying self-attention (Vaswani et al., 2017) along the height axis and the width axis of the entity table. Each computation along these axes is followed by a residual connection. The process for a pair (e_h, e_t) is as follows:

$$r_w^{(h,t)} = r_h^{(h,t)} + \sum_{p \in 1..n} \text{softmax}_p(q_{(h,t)}^T k_{(h,p)}) v_{(h,p)}$$
$$r_h^{(h,t)} = \mathbf{p}_{h,t} + \sum_{p \in 1..n} \text{softmax}_p(q_{(h,t)}^T k_{(p,t)}) v_{(p,t)} \qquad (21)$$

where the query $q_{(i,j)} = W_Q \mathbf{p}_{i,j}$, key $k_{(i,j)} = W_K \mathbf{p}_{i,j}$, and value $v_{(i,j)} = W_V \mathbf{p}_{i,j}$ are linear projections of the entity pair representation $\mathbf{p}$ at the position (i, j). The matrices W_Q, W_K, W_V are learnable weight parameters.

A.3.2 Objective Function

To enhance the robustness of the baseline DocRE system against the significant PN imbalance issue, we utilize Pairwise Moving-Threshold Loss combined with Entropy Minimization (Guo et al., 2023) as the primary objective, denoted as $\mathcal{L}_{pmt}^{h,t}$.

Pairwise Moving-Threshold Loss with Entropy Minimization ($\mathcal{L}_{pmt}^{h,t}$) For each entity pair (e_h, e_t), the relation set $\mathcal{R}$ is divided into a positive relation set $\mathcal{P}_{h,t}$ and a negative relation set $\mathcal{N}_{h,t}$. Given the prediction score f_r for a relation $r \in \mathcal{R}$ and the prediction score f_η for the NA class of (e_h, e_t), the probability that the relation C of (e_h, e_t), denoted as $P_{h,t}^r(r)$, is r given that C is either r or the NA class is computed as follows:

$$P_{h,t}^r(r) = \frac{\exp(f_r)}{\exp(f_r) + \exp(f_\eta)} \qquad (22)$$

Inversely, the probability of C is NA class is as:

$$P_{h,t}^{\text{NA}}(r) = 1 - P_{h,t}^r(r) = \frac{\exp(f_\eta)}{\exp(f_r) + \exp(f_\eta)} \qquad (23)$$

Then, the *pairwise moving-threshold* loss $\mathcal{L}_{pmt}^{h,t}$ is computed as follow:

$$
\begin{aligned}
\mathcal{L}_{pmt}^{h,t} &= -\log\Big(\prod_{r\in\mathcal{P}_{h,t}} P_{h,t}^r(r)\prod_{r\in\mathcal{N}_{h,t}}\big(1-P_{h,t}^r(r)\big)\Big) \\
&= -\sum_{r\in\mathcal{P}_{h,t}}\log P_{h,t}^r(r) - \sum_{r\in\mathcal{N}_{h,t}}\log P_{h,t}^{\mathsf{NA}}(r) \\
&= \sum_{r\in\mathcal{P}_{h,t}}\log(1+\exp(f_\eta - f_r)) \\
&\quad + \sum_{r\in\mathcal{N}_{h,t}}\log(1+\exp(f_r - f_\eta)).
\end{aligned}\tag{24}
$$

To further discriminate the margin in order to reduce the certainty of model prediction about positive and negative relations, (Guo et al., 2023) compute the information entropy for each pairwise probability distribution between relation r and the NA class for each (e_h, e_t):

$$
H_{h,t}(r) = -P_{h,t}^r(r)\log P_{h,t}^r(r) - P_{h,t}^{\mathsf{NA}}(r)\log P_{h,t}^{\mathsf{NA}}(r).\tag{25}
$$

Finally, the pairwise moving-threshold loss with entropy minimization is as follows:

$$
\mathcal{L}_{em}^{h,t} = \frac{1}{\gamma_1}\sum_{r\in\mathcal{P}_{h,t}}H_{h,t}(r) + \frac{1}{\gamma_2}\sum_{r\in\mathcal{N}_{h,t}}H_{h,t}(r),\tag{26}
$$

$$
\mathcal{L}_1 = \sum_{(h,t)\in\mathcal{B}}\mathcal{L}_{pmt}^{h,t} + \mathcal{L}_{em}^{h,t},\tag{27}
$$

where $\mathcal{B}$ refers to a training batch, and $\gamma_1 = \{1, |\mathcal{P}_{h,t}|\}$ and $\gamma_2 = \{1, |\mathcal{N}_{h,t}|\}$ are hyperparameters weighting the effect of entropy minimization.

Supervised Contrastive Learning $(\mathcal{L}_{scl}^{h,t})$ We further incorporate the Supervised Contrastive Learning loss (Khosla et al., 2020) to enhance the performance the DocRE model:

$$
\mathcal{L}_{scl}^{h,t} = -\log\left\{\frac{1}{|\mathcal{S}_{h,t}|}\sum_{p\in\mathcal{S}_{h,t}}\frac{\exp(\boldsymbol{x}_{h,t}\cdot\boldsymbol{x}_p/\tau)}{\sum_{d\in\mathcal{B},d\neq(h,t)}\exp(\boldsymbol{x}_{h,t}\cdot\boldsymbol{x}_d/\tau)}\right\},\tag{28}
$$

where $\mathcal{B}$ is the set of entity pairs including (h,t) in the current batch. $\mathcal{S}_{h,t}\subseteq\mathcal{B}$ is the set in which each entity pair $p=(h',t')$ has at least one common positive relation with $(h,t)\notin\mathcal{S}_{h,t}$, and p is termed a *positive* example of (h,t). The *negative* examples of (h,t) are the remaining examples in $\mathcal{B}$. The operator $(\cdot)$ refers to the dot product, and $\tau\in\mathbb{R}^+$ is a temperature parameter.

Final Objective Functions $(\mathcal{L}_{\textbf{DocRE}})$ Additionally, we apply the Supervised Contrastive Learning

loss, denoted as $\mathcal{L}_{scl}^{h,t}$, to emphasize the distinctions between the embeddings of entity pairs that express different relations. Ultimately, the baseline DocRE model aims to minimize the following loss function:

$$
\mathcal{L}_{\mathrm{DocRE}} = \mathcal{L}_{pmt}^{h,t} + \lambda\times\mathcal{L}_{scl}^{h,t}\tag{29}
$$

where λ is the hyper-paremeter to control the strength of $\mathcal{L}_{scl}^{h,t}$.

A.4 Derivation of Equation (9)

First, Eq. (11) is reformulated using Eq. (8):

$$
\begin{aligned}
\mathrm{ELBO}(\boldsymbol{\theta},\boldsymbol{\phi},\psi;\mathbf{p}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{p}}|\mathbf{p})}[\log(f_{\boldsymbol{\theta}}(\mathbf{p}|\mathbf{z}_{\mathbf{p}}))] \\
&\quad + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{p}}|\mathbf{p})}[-\log(q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{p}}|\mathbf{p}))] \\
&\quad + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{p}}|\mathbf{p})}[\log(p_\psi(\mathbf{z}_{\mathbf{p}}))]
\end{aligned}\tag{30}
$$

where p_ψ denotes the diffusion model, $p_{\boldsymbol{\theta}}$ is the VAE decoder, and $q_{\boldsymbol{\phi}}$ represents the VAE encoder.

Next, the final term in Eq. (30) is expanded using Eq. (16), yielding the following:

$$
\begin{aligned}
&\mathbb{E}_{q_{\boldsymbol{\phi}}}[\log(f_{\boldsymbol{\theta}}(\mathbf{p}|\mathbf{z}_{\mathbf{p}_0})) - \log(q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{p}_0}|\mathbf{p})) + \\
&\qquad \mathbb{E}_{\mathbf{z}_{\mathbf{p}_{0:T}}\sim q_{\boldsymbol{\phi}}}\left[\log\frac{p_\psi(\mathbf{z}_{\mathbf{p}_{0:T}})}{q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{p}_{1:T}}|\mathbf{z}_{\mathbf{p}_0})}\right]
\end{aligned}\tag{31}
$$

$$
\begin{aligned}
&\leq \mathbb{E}_{q_{\boldsymbol{\phi}}}[\log(f_{\boldsymbol{\theta}}(\mathbf{p}|\mathbf{z}_{\mathbf{p}_0})) - \log(q_{\boldsymbol{\phi}}(\mathbf{z}_{\mathbf{p}_0}|\mathbf{p}) \\
&\quad + \log(p_\psi(\mathbf{z}_{\mathbf{p}_0}))] \leq \log(f_{\boldsymbol{\theta}}(\mathbf{p}|\mathbf{z}_{\mathbf{p}_0}))
\end{aligned}\tag{32}
$$

where $\mathbf{z}_{\mathbf{p}_0}$ corresponds to the latent space representation generated by the EP-VAE module.

Eqs. (31) and (32) establish a valid ELBO formulation. Maximizing Eq. (31) is equivalent to minimizing the VaeDiff module's objective function. Additionally, maximizing the final term in Eq. (31) aligns with minimizing the Diffusion Prior's objective function. By substituting the last term of Eq. (31) with the Diffusion Prior loss, we derive Eq. (9).