Pinpointing Diffusion Grid Noise to Enhance Aspect Sentiment Quad Prediction

Linan Zhu[1], Xiangfan Chen[1], Xiaolei Guo[1],
Chenwei Zhang[2], Zhechao Zhu[1], Zehai Zhou[1], Xiangjie Kong[1]*
[1]Zhejiang University of Technology
[2]The University of Hong Kong
zln@zjut.edu.cn, xiangfan_chen@outlook.com, xjkong@ieee.org

Abstract

Aspect sentiment quad prediction (ASQP) has garnered significant attention in aspect-based sentiment analysis (ABSA). Current ASQP research primarily relies on pre-trained generative language models to produce templated sequences, often complemented by grid-based auxiliary methods. Despite these efforts, the persistent challenge of generation instability remains unresolved and the effectiveness of grid methods remains underexplored in current studies. To this end, we introduce **G**rid Noise **D**iffusion **P**inpoint Network (**GDP**), a T5-based generative model aiming to tackle the issue of generation instability. The model consists of three novel modules, including Diffusion Vague Learning (DVL) to facilitate effective model learning and enhance overall robustness; Consistency Likelihood Learning (CLL) to discern the characteristics and commonalities of sentiment elements and thus reduce the impact of distributed noise; and GDP-FOR, a novel generation template, to enable models to generate outputs in a more natural way. Extensive experiments on four datasets demonstrate the remarkable effectiveness of our approach in addressing ASQP tasks.[1]

1 Introduction

Aspect sentiment quad prediction (ASQP) has recently attracted widespread attention in the field of aspect-based sentiment analysis (ABSA), a fine-grained sentiment analysis task that aims to extract more comprehensive sentiment elements that include (1) *aspect terms (at)*; (2) *opinion terms (ot)*; (3) *aspect categories (ac)*; (4) *sentiment polarity (sp)*. An instance is shown in the upper half of Figure 1.

ASQP subtasks usually entail identifying *at* and excavating its corresponding *ac*, then establishing connections between them by incorporating supporting *ot* and/or *sp*. However, more than 30% of sentiment expressions manifest implicitly (Peper and Wang, 2022; Cai et al., 2021) in practice, posing significant challenges in accurately extracting quadruples from sentences containing implicit *at* and/or *ot*.

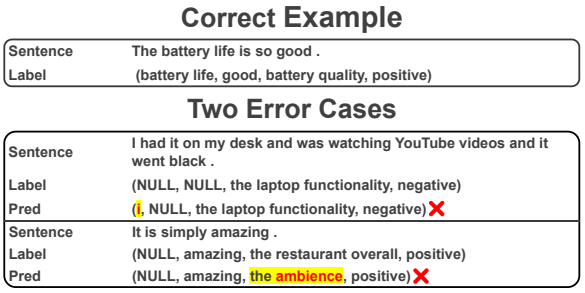


Figure 1: The results of Label and Pred are presented in the order of (*at, ot, ac, sp*), with items containing prediction errors highlighted in red font text.

Owing to its extensive applicability across various scenarios, considerable efforts have been devoted to ABSA (Pontiki et al., 2016; Zhang et al., 2022a). Presently, there are two predominant methodologies. (1) Pipeline approaches (Zhang et al., 2022a) render multitasking learning more intuitive by dedicating each module to solving specific tasks. However, this method often overlooks inter-element relationships, rendering it susceptible to the cascading effects of error propagation (Wu et al., 2020). (2) Generative approaches (Zhang et al., 2022a) enable end-to-end solutions for ASQP, thereby mitigating potential error propagation encountered in pipeline-based approaches. By learning to generate sentiment elements in the form of natural language, these approaches harness the semantics of sentiment elements to the fullest, specifically addressing the challenges associated with easily omitted *at* and *ot* elements.

While the aforementioned methodologies offer

*Corresponding author
[1]Code for our method is available at : https://github.com/ch11en/GDP_

Findings of the Association for Computational Linguistics: ACL 2024, pages 3717–3726

valuable insights into addressing ASQP tasks, they fall short in generating stable outputs, including errors in outputting semantically similar words, inclusion of unfriendly implicit words, and output duplication, as illustrated in the lower half of Figure 1. Xu et al. (2021) introduced a span-based method to address these issues by serializing sentiment-related "word sequences" and rating them. However, this approach is prone to overlooking scattered or implicit sentiment information. Wu et al. (2020) proposed a table-filling approach, but this method diminishes the model's sensitivity to the interplay among elements. Furthermore, Zhang et al. (2022b) argue that independently assigning each label in the table-filling method may lead to discrepancies, potentially resulting in incorrect output predictions.

Therefore, we take notice of the diffusion models (Ho et al., 2020; Li et al., 2022) that were employed to address noise in signal propagation. The training process of the diffusion model involves two stages: the forward noise addition process and the reverse denoising process. The model is based on a neural network, which has memory and denoising capabilities and can effectively learn the characteristics of the datasets. Drawing inspiration from these models, we introduce **G**rid Noise **D**iffusion **P**inpoint Network (**GDP**), a T5-based generative model aiming to tackle the issue of generation instability. Within our model, we propose the Diffusion Vague Learning (DVL) mechanism to emulate the forward process to construct vague grids and utilize a golden grid to guide the backward convergence process. Furthermore, we introduce a novel loss calculation target called Consistency Likelihood Learning (CLL) to consolidate identical sentiment elements and thereby reduce each elements distribution noise generated during training process. Finally, we design a novel output template named GDP-FOR to enhance the naturalness of the generated results. The contributions of this paper are outlined as follows:

- We propose GDP to mitigate the inherent mistakes of pre-trained language models by incorporating DVL into the T5 model. To the best of our knowledge, this is the first attempt to combine the diffusion concept into the ASQP task.

- We propose CLL to bring different sentiment elements within the same quadruple closer together. This objective effectively improves the model's ability to discriminate complex input elements and reduce vague noise, thereby assisting the model in extracting quadruples.

- We design a new generation template called GDP-FOR, which takes into account both grammar and human intuition, in order to enable the model to produce results in a more natural manner.

- We conducted extensive experiments on Res15, Res16, Laptop, and Restaurant datasets. The results demonstrate that GDP can offer superior performance over baselines, which indicates the universal effectiveness of our model.

2 ASQP Task Definition

We adhere to the definitions established in previous works focused on generation-based methods (Hu et al., 2022; Zhang et al., 2021a; Cai et al., 2021; Hu et al., 2023). The ASQP task aims to predict all aspect-level quadruplets $\mathbf{Q_1}$, $\mathbf{Q_2}$, ..., $\mathbf{Q_n}$ in the source sentence $\mathbf{S}$, where $\mathbf{Q_i} = (at_i, ac_i, ot_i, sp_i)$. Each component represents the aspect term, aspect category, opinion term, and sentiment polarity. Quadruples will be marked as *implicit* when lacking clear *at* or *ot* (Peper and Wang, 2022).

3 Methodology

In this section, we will introduce the constituent elements of GDP: (1) Diffusion Vague Learning (DVL), a noise mitigation module; (2) Consistency Likelihood Learning (CLL), a novel objective for calculating sentiment distribution that aids in reducing noise generated by various sentiment elements; (3) GDP-FOR, a rational generation format that is more suitable for realistic scenarios. These approaches are specifically designed to enhance the model's intentional generation capabilities and mitigate the impact of implicit expressions noise and other noise. The overall structure is depicted in Figure 2.

3.1 Diffusion Vague Learning

The distribution of sentiment elements and the mapping vector space generated by a semantically similar sentence are often closely aligned (Peper and Wang, 2022), posing challenges in differentiating noise from the actual sentence and extracting the

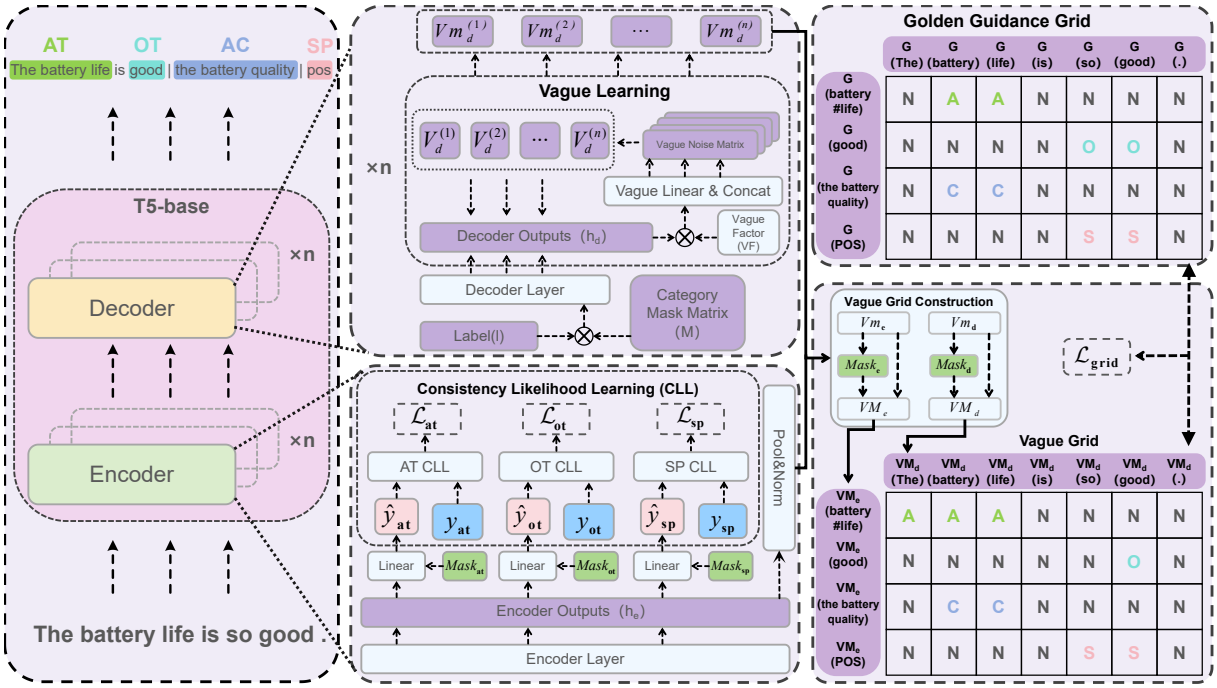


Figure 2: Overall structure of GDP model.

correct quadruple. To address this concern, we introduce DVL to acquire the desired samples. Each sample is derived through multiple self-attention mechanisms within the enhanced transformer architecture (Vaswani et al., 2017).

Vague Matrix Learning As can be seen in the the upper middle part of Figure 2, The states h_e obtained from the encoder layer undergo a linear layer to construct the vertical axis of the vague grid.

$$Vm_e = Softmax(Norm(Pool(Linear(h_e)))) \quad (1)$$

The label l and the aspect category mask matrix M are combined and subsequently sent to the DecLayer (Decoder layer) for training, where pad_{id} represents pad token ID that set to -100 and t denotes the iteration step.

$$h_d^{(t)} = DecLayer(l \cdot M^{(t)} + (1 - M^{(t)}) \times (pad_{id})) \quad (2)$$

Building upon the conception of the Diffusion model (Ho et al., 2020) and Diffusion-LM (Li et al., 2022), we insert a trace of vague factors $VF^{(t)}$, which follows a Gaussian distribution, into $h_d^{(t)}$ at each time step t to simulate the forward diffusion process.

$$V_d^{(t)} = h_d^{(t)} \cdot VF^{(t)} \quad (3)$$

Subsequently, the outputs $V_d^{(t)}$ are concatenated to obtain a vague matrix Vm_d. Afterward, we feed

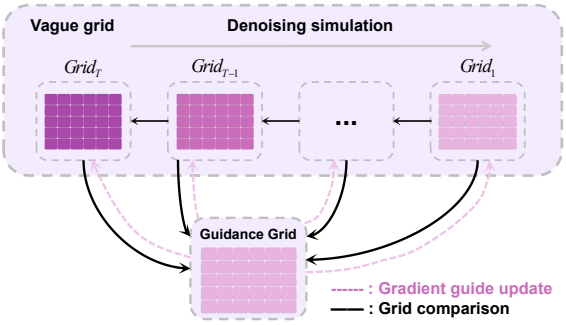


Figure 3: Denoising simulation. This part is an explanation of the third part in the overall structure diagram of the model of Figure 2

Vm into a linear layer to produce the final result after softmax operations:

$$Vm_d = Softmax(Linear(Concat(V_d^{(1)}, ..., V_d^{(t)}))) \quad (4)$$

Vague Grid Loss Objective The outputs from the encoder Vm_e and decoder Vm_d are combined through concatenation. Subsequently, these merged outputs are fed into the grid model for integration, forming a comprehensive grid G for predictions.

$$G^{(t)} = Grid\ Linear \left(\begin{bmatrix} [Vm_e^{(1)}, Vm_d^{(1)}] & \dots & [Vm_e^{(1)}, Vm_d^{(m)}] \\ \vdots & \ddots & \vdots \\ [Vm_e^{(n)}, Vm_d^{(1)}] & \dots & [Vm_e^{(m)}, Vm_d^{(m)}] \end{bmatrix} \right) \quad (5)$$

As depicted in the right portion of Figure 2, a reference grid, known as the "Golden Guidance Grid", is utilized as an instructor to steer the backward convergence process. The iterative process is

depicted in Figure 3.

$$\mathcal{L}_{grid} = -\frac{1}{T}\sum_{t\in T} G_{golden\ grid} \cdot \log(G^{(t)}) \quad (6)$$

3.2 Sentiment Distribution Learning

Consistency Likelihood Learning CLL seeks to understand the similarity in consistency distribution for each sentiment element to reduce the probability of noise generation. As illustrated in Figure 2, we observe the likelihood of each sentiment element aligning with the ground-truth sentiment label $\hat{y}_{at}$, $\hat{y}_{ot}$, $\hat{y}_{sp}$ (abbreviated as $\hat{y}_{(at,ot,sp)}$, the following *y* and *Mask* are the same).

$$y^{(t)}_{(at,ot,sp)} = Linear(H^{(t)}_e \cdot Element\ Mask_{(at,ot,sp)}) \quad (7)$$

For each correct sentiment element, we need to bring it closer to the corresponding correct sentiment label $\hat{y}_{(at,ot,sp)}$. For each ambiguous sample elements, we differentiate error samples to the greatest extent possible. In this way, both the correct and incorrect elements can be considered simultaneously.

$$\mathcal{L}_{(at,ot,sp)} = -\sum_{t\in T} y^{(t)}_{(at,ot,sp)} \log(\hat{y}_{(at,ot,sp)}) \quad (8)$$

Joint Distribution Learning Objective Joint distribution learning considers the distribution of each learned sentiment element and the influence of the vague noise factor. The ultimate training objective is to concurrently integrate the four aforementioned losses:

$$\mathcal{L}_{total} = \mathcal{L}_{grid} + \mathcal{L}_{at} + \mathcal{L}_{ot} + \mathcal{L}_{sp} \quad (9)$$

3.3 Structured Generation Format

Previous Structured Generation Formulation In a previous work, Zhang et al. (2021a) linearize the $\mathbf{Q_i} = (at_i, ac_i, ot_i, sp_i)$ into an output format $P(\mathbf{Q_i})$ and utilize it as the generation goal:

$$P(\mathbf{Q_i}) = P_{ac}(ac_i) \text{ is } P_{sp}(sp_i) \text{ because } P_{at}(at_i) \text{ is } P_{ot}(ot_i) \quad (10)$$

The mapping function is as follows: (1) If *at* and/or *ot* are explicitly stated, $P_{at} = at$ and/or $P_{ot} = ot$; otherwise, P_{at} and/or P_{ot} is set to *NULL*; (2) *ac* will be converted into the pre-defined mapping rule $\mathbf{C}$, such as $P_{ac} = \{the\ support\ quality\}$ for $ac = \{SUPPORT\#QUALITY\}$; (3) *sp* will be mapped into sentiment semantic words {great, ok, bad} for clearer expression. The upper left corner of Figure 2 shows the quadruple elements.

Datasets	Laptop	Restaurant
#Categories	121	13
#Sentences	4076	2286
#EAEO Quads	3269 (56.8%)	2429 (66.40%)
#EAIO Quads	1237 (21.5%)	350 (9.57%)
#IAEO Quads	910 (15.8%)	530 (14.5%)
#IAIO Quads	342 (5.94%)	349 (9.54%)

Table 1: ACOS dataset statistics.

Datasets		Res15	Res16	Restaurant	Laptop
Train	#Sent	834	1264	2934	1530
	#Quad	1354	1989	4172	2484
Dev	#Sent	209	316	326	171
	#Quad	347	507	440	261
Test	#Sent	537	544	816	583
	#Quad	795	799	1161	916
Quad/Sent ratio (%)		1.58	1.55	1.42	1.60

Table 2: Overall statistics.

If a sentence contains multiple quadruples, a special separator token [SSEP] will be inserted to create segmentation. In the end, we obtain the output:

$$P(\mathbf{Q_1})\ [SSEP]\ P(\mathbf{Q_2})\ ...\ [SSEP]\ P(\mathbf{Q_n}) \quad (11)$$

Hu et al. (2022) utilize special tokens to separate individual sentiment elements and demonstrate that the effectiveness of a template order varies across diverse datasets. An example of one of their templates is provided below:

$$[AT]\ x_{at}\ [OT]\ x_{ot}\ [AC]\ x_{ac}\ [SP]\ x_{sp} \quad (12)$$

GDP Structured Generation Format In this work, we design a more rational generation format named GDP-FOR for the GDP model. Specifically, we give *at* the highest priority, followed by *ot* to ensure logical coherence and alignment with human intuition. Subsequently, due to the strong affiliation among *ac*, *at*, and *ot*, *ac* is placed third in order. Finally, *sp* is required to encompass all fine-grained sentiment elements, so it is positioned at the end of the template. The procedure is outlined as follows:

$$\text{The } P_{at}(at) \text{ is } P_{ot}(ot) \mid P_{ac}(ac) \mid P_{sp}(sp) \quad (13)$$

4 Experimental Setups

In this section we introduce the detailed experimental setups on which our experiments rely.

Datasets We conduct experiments on four datasets: Res15, Res16, Restaurant, and Laptop.

The first two datasets, proposed by (Zhang et al., 2021a), are classic ASQP task datasets, while the last two, proposed by (Cai et al., 2021), include numerous implicit words. Detailed ACOS statistics are provided in Table 1, and the overall statistics are presented in Table 2.

Evaluation Metrics The F1 scores used in our method are the primary evaluation metric, and corresponding precision and recall scores will also be reported. The correctness of a predicted quad is affirmed when all the predicted elements precisely match the gold labels.

Implementation Details The T5-BASE model (Raffel et al., 2020) serves as the foundation for implementing the GDP model, employing a traditional Transformer encoder-decoder architecture. The experiment is conducted using PyTorch version 1.13.1 and executed on a single Nvidia RTX 3090 GPU. During the training phase, we set the batch size, training epochs, and optimizer parameters to 8, 20, and Adam (Kingma and Ba, 2014), respectively. The learning rate was fine-tuned on various datasets to optimize performance. In the inference phase, a beam search with a value of 5 is employed for generating output sentences. Although we experimented with setting the beam search from 2 to 10, the results indicated that setting it to 5 yields the best outcomes. Ultimately, we present the results based on the averaged scores obtained from 5 runs with different random seed initializations.

Comparison Method We meticulously compare the GDP model against a selection of robust baselines, encompassing both non-generation and generation methods. (1) **Pipeline Methods**: Double-Propagation-ACOS (Cai et al., 2021), JET-ACOS (Xu et al., 2020), TAS-BERT-ACOS (Wan et al., 2020), Extract-Classify-ACOS (Wang et al., 2017), TASO-BERT-Linear-CRF (Zhang et al., 2021a); (2) **Generation Methods**: GAS (Zhang et al., 2021b), Paraphrase (Zhang et al., 2021a), Seq2Path (Mao et al., 2022), GEN-SCL-NAT (Peper and Wang, 2022), Special_Symbols (Hu et al., 2022), Special_Symbols+UAUL (Hu et al., 2023), OTG (Bao et al., 2022), MVP (Gou et al., 2023), DLO (Hu et al., 2022), DLO+UAUL (Hu et al., 2023).

5 Experimental Results

5.1 Overall Results

The experimental results of the model on the ACOS and ASQP datasets are presented in Table 3 and 4. In terms of the ACOS datasets, GDP demonstrates superior performance in F1 score compared to other methods. A particularly notable observation is the comparison with the robust baseline OTG, where the GDP model showcases absolute F1 score enhancements of 1.38% on the Restaurant dataset and 0.44% on the Laptop dataset. GDP also exhibits commendable performance on the two ASQP datasets. While the GDP model may exhibit suboptimal results on Res15, it attains state-of-the-art results on Res16 dataset. Importantly, when juxtaposed with a robust baseline like DLO+UAUL, the GDP model achieves noteworthy absolute F1 score improvements of 1.11%.

Pipeline methods, placed above the dashed line in both tables, have shown deficiencies in terms of all indicators. These challenges could potentially arise from error accumulation across multiple subtasks. Among the methods below the dashed line in both tables, the inconspicuous performance of generative methods may be attributed to two factors: limited sensitivity to noise factors and weak denoising ability in model generation. Therefore, we posit that the robust performance of the GDP model primarily stems from its improved ability to capture noise in multi-scale local and global features through a meticulously designed forward diffusion and reverse guidance process, along with a carefully crafted output template. Additionally, the model introduces a novel sentiment distribution computation objective, enabling it to obtain a clear distribution of consistent sentiment labels within the same quadruple for better extraction of both explicit and implicit aspect sentiment quadruples in sentences.

5.2 Ablation Study

To assess the effectiveness of individual components, we conducted a comprehensive ablation study on the GDP model by excluding various components of the technique. The ablation study outcomes on the ACOS and ASQP datasets are presented in Table 5 and 6, respectively. Specifically, the '✗' and '✓'in the first three columns respectively indicate whether a module is included or not. It can be observed that by removing various components, performance on the four datasets consistently decreases (average decrease of 1.91%).

DVL Ablation Our initial goal in building the DVL module is to assist the model in recognizing both explicit and implicit noise within the

Methods	Laptop			Restaurant		
	Pre %	Rec %	F1 %	Pre %	Rec %	F1 %
Double-Propagation-ACOS	13.04 ▼33.8%	0.57 ▼43.63%	8.00 ▼37.48%	34.67 ▼30.04%	15.08 ▼48.63%	21.04 ▼43.17%
JET-ACOS	44.52 ▼2.32%	16.25 ▼27.95%	23.81 ▼21.67%	59.81 ▼4.9%	28.94 ▼34.77%	39.01 ▼25.2%
TAS-BERT-ACOS	**47.15** ▲0.31%	19.22 ▼24.98%	27.31 ▼18.17%	26.29 ▼38.42%	46.29 ▼17.42%	33.53 ▼30.68%
Extract-Classify-ACOS	45.56 ▼1.28%	29.48 ▼14.72%	35.80 ▼9.68%	38.54 ▼26.17%	52.96 ▼10.75%	44.61 ▼19.6%
GAS	41.60 ▼5.24%	42.75 ▼1.45%	42.17 ▼3.31%	60.69 ▼4.02%	58.52 ▼5.19%	59.59 ▼4.62%
Paraphrase	41.77 ▼5.07%	42.56 ▼1.64%	43.34 ▼2.14%	58.98 ▼5.73%	59.11 ▼4.60%	59.04 ▼5.17%
Seq2Path	-	-	42.97 ▼2.51%	-	-	58.41 ▼5.8%
GEN-SCL-NAT	-	-	45.16 ▼0.32%	-	-	62.62 ▼1.59%
MVP	-	-	43.92 ▼1.56%	-	-	61.54 ▼2.67%
Special_Symbols	43.58 ▼3.26%	42.72 ▼1.48%	43.15 ▼2.33%	59.98 ▼4.73%	58.40 ▼5.31%	59.18 ▼5.03%
OTG	46.11 ▼0.73%	**44.79** ▲0.59%	45.44 ▼0.04%	63.96 ▼0.75%	61.74 ▼1.97%	62.83 ▼1.38%
GDP (ours)	46.84	44.20	**45.48**	**64.71**	**63.71**	**64.21**

Table 3: Comparison results on ACOS datasets. The best results are highlighted in bold, and the suboptimal results underlined. All comparative model data were obtained from the corresponding papers.

Methods	Res15			Res16		
	Pre %	Rec %	F1 %	Pre %	Rec %	F1 %
TASO-BERT-Linear	41.86 ▼7.34%	26.50 ▼8.45%	32.46 ▼17.29%	49.73 ▼11.43%	40.70 ▼21.38%	44.77 ▼16.84%
TASO-BERT-CRF	44.24 ▼4.96%	28.66 ▼21.65%	34.78 ▼14.97%	48.65 ▼12.51%	39.68 ▼22.40%	43.71 ▼17.90%
Extract-Classify-ACOS	35.64 ▼13.56%	37.25 ▼13.06%	36.42 ▼13.33%	38.40 ▼22.76%	50.93 ▼11.15%	43.77 ▼17.84%
GAS	45.31 ▼3.89%	46.70 ▼3.61%	45.98 ▼3.77%	54.54 ▼6.62%	57.62 ▼4.46%	56.04 ▼5.57%
Paraphrase	46.16 ▼3.04%	47.72 ▼2.59%	46.93 ▼2.82%	56.63 ▼4.53%	59.30 ▼2.78%	57.93 ▼3.68%
Special_Symbols+UAUL	49.12 ▼0.08%	50.39 ▲0.08%	**49.75**	59.24 ▼1.92%	61.75 ▼0.33%	60.47 ▼1.14%
DLO+UAUL	48.03 ▼1.17%	50.54 ▲0.23%	49.26 ▼0.49%	59.02 ▼2.14%	62.05 ▼0.03%	60.50 ▼1.11%
GDP (ours)	**49.20**	50.31	**49.75**	**61.16**	**62.08**	**61.61**

Table 4: Comparison results on ASQP datasets. The best results are highlighted in bold, and the suboptimal results underlined. All comparative model data were obtained from the corresponding papers.

data. Therefore, we conducted ablation experiments by removing its constituent part to demonstrate whether the DVL module is valid. The results reveal the most significant degradation across all datasets (an average decrease of 1.31% in the ACOS datasets and 1.06% in the ASQP datasets), substantiating its ability to aid the model in recognizing noise and contributing to more accurate quadruplet extraction. Notably, performance degradation on the ASQP datasets is not as pronounced as on the ACOS datasets. We posit that the abundance of implicit words in the ACOS dataset sparked the model's interest in learning these nuances, thereby assisting in generating "more pure" template sentences.

CLL Ablation We contend that the elimination of the CLL component, which involves processing the original sentiment elements directly rather than the clustered sentiment elements, would lead to a reduction in the model's sensitivity to the distribution of sentiment elements. The results from the two tables clearly indicate a consistent decline in performance across all datasets (an average decline of 1.29% in the ACOS datasets and 0.85% in ASQP datasets) upon removing these components. This underscores the efficacy of the CLL module in helping the model attain deeper and more precise insights. CLL encourages the model to capture subtle connections among various sentiment elements by clustering similar elements and distancing different elements, effectively narrowing the gap between correct words, similar words, and implicit elements. Ultimately, this aids the model in achieving excellent performance.

GDP-FOR Ablation To validate the effectiveness of the GDP-FOR components, we replace GDP-FOR with the original generation template (10). As depicted in Tables 5 and 6, excluding GDP-FOR has a noticeable negative impact on overall performance (an average decrease of 1.01% in the ACOS datasets and 0.74% in the ASQP datasets). Particularly noteworthy is the substantial impact on the Laptop dataset (45.48% $\rightarrow$ 44.04%), distinguished by a higher number of *ac* (121 vs. 13 for the Laptop and Restaurant datasets), an increased volume of sentences, and more intricate implicit representations than the Restaurant datasets. We posit that another reason for the diminished performance upon removing GDP-FOR is the redundancy present in the original template, which poses challenges for accurate generation by the model. This confirms our belief that a meticulously de-

DVL	CLL	GDP-FOR	Restaurant	Laptop	Avg
✓	✓	✓	64.21	45.48	-
✗	✓	✓	63.13▼1.08%	43.94▼1.54%	▼1.31%
✓	✗	✓	63.17▼1.04%	43.95▼1.53%	▼1.29%
✓	✓	✗	63.48▼0.73%	44.04▼1.44%	▼1.01%
✗	✗	✗	61.94▼2.27%	43.57▼2.14%	▼2.09%

Table 5: Ablation study results on ACOS datasets. We use F1 score as the main criterion for consideration.

DVL	CLL	GDP-FOR	Res15	Res16	Avg
✓	✓	✓	49.75	61.61	-
✗	✓	✓	48.72▼1.03%	60.53▼1.08%	▼1.06%
✓	✗	✓	48.91▼0.84%	60.76▼0.85%	▼0.85%
✓	✓	✗	49.02▼0.73%	60.87▼0.74%	▼0.74%
✗	✗	✗	48.12▼1.63%	59.78▼1.83%	▼1.73%

Table 6: Ablation study results on ASQP datasets. We use F1 score as the main criterion for consideration.

signed output template is crucial for handling multi-quadruple examples.

In summary, our complete GDP method consistently delivers robust performance, with its components synergistically addressing the intricate challenges inherent in the ACOS and ASQP tasks.

5.3 Analysis of Vague Grid Weight

As illustrated in Figure 4a, we present the results of the GDP model across different grid weight configurations on four datasets to analyze the impact of varying vague grid weights for the loss function on performance. It is observed that as λ_{grid} gradually increases from 0 to 1, the F1 values for all datasets initially rise then fluctuate and decline, which confirms that simulating the iterative process of diffusion models using a vague grid can effectively improve sensitivity toward noise factors generated during training. Moreover, we observe that the Res15 and Restaurant datasets are less affected by changes in grid weights. We attribute this difference to the sent/quad ratio in each dataset. The higher the proportion, the more quadruples a sentence contains, which invisibly creates difficulties for model generation.

5.4 Analysis of Iterative Steps in Diffusion Vague Learning

In this subsection, we conducted ablation experiments on the number of iteration steps for adding noise during the DVL stage of the model. Figure 4b shows our experimental results on the four datasets. As the number of time steps progresses from 0 to 20, the F1 performance of model transitions from subpar to satisfactory, eventually declining, which proves that the number of time steps in simulating diffusion processes has a significant impact on model performance. A plausible interpretation suggests that when the number of time steps is low, the noise level is insufficient for the model to capture obvious features of sentiment elements. Conversely, with a large number of time steps, the model struggles to accommodate all the noisy data, highlighting the need for further enhancements in our diffusion simulation model.

5.5 Error Analysis and Case Study

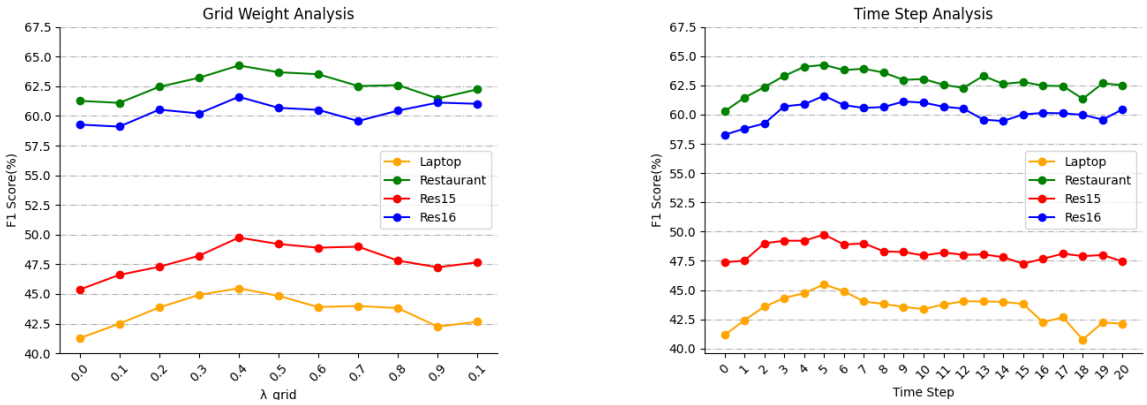
We conducted an error analysis on our model with the aim of gaining a comprehensive understanding of its prediction behavior and identifying the reasons for errors. Randomly selecting 100 sentences from the Dev set of each dataset, we utilized the trained model to generate predictions. The generation results are illustrated in Figure 5.

In the first example, the actual *ot* in the label is 'NULL'. However, our model implicitly characterizes the *ot* as 'expected'. Similarly, in the second case, the output is 'i' in the position of the *at*, whereas the expected label is 'NULL'. Although the *at* is explicit, our approach incorrectly predicts the *at*. We believe there are two possible reasons for this phenomenon: (1) the model may not have learned sufficiently rich semantic representations; (2) it may have learned excessive noise information in vague matrix learning. While GDP has consistently achieved performance improvements across various generative methods, particularly in addressing issues related to implicit information sensitivity, further research is needed to assist models in distinguishing implicit elements.

6 Related Work

6.1 Aspect-Level Sentiment Analysis

Early investigations into aspect-based sentiment analysis (ABSA) primarily focused on individual elements, such as aspect term extraction (ATE) (Liu et al., 2015; Xu et al., 2018), aspect category detection (ACD) (Zhou et al., 2015; Tulkens and van Cranenburgh, 2020), opinion term extraction (OTE) (Li et al., 2018; Mensah et al., 2021), and aspect sentiment classification (ASC) (Wang et al., 2021; Zhou et al., 2021). At present, researchers are progressively placing greater emphasis on extracting compound sentiment elements (Zhu et al., 2023). Scholars represented by (Zhao et al., 2020; Liang et al., 2020; Liu et al., 2021) have refined the single element extraction and proposed some branch direction such as aspect-opinion pair extrac-



(a) Evaluation results of vague grid weight analysis

(b) Evaluation results of time step analysis

Figure 4: Analysis of the impact of module weights on model performance.

Error Cases Predicted By GDP

Sentence	this computer was 10 fold what i had expected.
Label	(computer, NULL, the laptop overall, positive)
Pred	(computer, expected, the laptop overall, positive) ✗
Sentence	i have been going back again and again.
Label	(NULL, NULL, the restaurant overall, positive)
Pred	(i, NULL, the restaurant overall, positive) ✗

Figure 5: GDP error case display.

tion (AOPE), End-to-End ABSA (E2E-ABSA), and aspect category sentiment analysis (ACSA). Peng et al. (2020) focus on aspect sentiment triplets.

Recently, Aspect Sentiment Quadruple Extraction (ASQP), a new research direction aiming to process whole sentiment elements, has drawn much attention. To implement this approach, researchers have introduced a pipeline method (Cai et al., 2021) and a generation-based method (Zhang et al., 2021a). Numerous experiments have shown that using generative methods for ASQP tasks is more effective, which is currently the mainstream approach because of its simplicity and end-to-end manner of generation. In subsequent improvement research, Hu et al. (2022) modified the output sequence of the model to demonstrate the effectiveness of element arrangement for generative ASQP tasks. Peper and Wang (2022) introduced contrastive learning to obtain representations of explicit and implicit examples to assist learning, and also achieved excellent results.

6.2 Concept of Diffusion Model

Diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020) were originally proposed in the field of computer vision to address the noise problem; these were originally latent variable models designed for continuous data domains. The model training process can be divided into two steps: the forward noise addition process and the reverse denoising process.

The forward process originates from data $x_0 \sim q(x)$. The model adds the noise corresponding to time step t and obtains output x_t according to x_{t-1}. At step T (the final time step) to obtain x_T, the data is transformed into an invisible noise distribution. In the reverse process, according to the given condition x_t (t decrements from T to 0), Bayes' theorem is used to determine x_{t-1}. As a result, the target sentence or image can be generated by iteratively sampling noise.

In contrast to the above works on ASQP, we borrowed the idea of the forward noise addition process and reverse denoising process from diffusion models to handle the noise during the training process.

7 Conclusion

In this article, we note that previous studies on ASQP have exclusively focused on essential content for model generation, overlooking the incorporation of various noise during the training process. In this research, we introduce a novel denoising approach called Diffusion Vague Learning (DVL) and seamlessly integrate it into the T5 model, thus establishing a GDP model. Simultaneously, we incorporate Consistency Likelihood Learning (CLL) to address the problem of sentiment element fragmentation during model training, effectively reducing the generation of noise. Finally, we meticulously design an output template for the GDP model that

aligns more closely with human intuition and grammatical rules. Extensive experiments demonstrate that the GDP method achieves cutting-edge results.

8 Limitations

Our work represents the pioneering exploration of integrating generative ASQP tasks with diffusion concepts. Despite achieving state-of-the-art performance, our approach grapples with certain limitations:

Firstly, GDP still encounters challenges when dealing with robust implicit works noise as well as other noise. Instances highlighted in the error analysis (refer to §5.5) underscore that complex cases demand a more profound semantic understanding. While GDP exhibits substantial advancements in the generation paradigm, determining the most effective type of noise treatment remains a challenge.

Secondly, the grid diffusion method necessitates the construction of a two-dimensional table representation. Consequently, the size of the table representation is notably larger than that of the sequence representation. Hence, our method consumes more training memory than alternative approaches.

We firmly believe that addressing the aforementioned limitations can lead to further enhancements in the model.

Acknowledgments This work was supported by the National Natural Science Foundation of China (No. 62176234, 62072409).

References

Xiaoyi Bao, Zhongqing Wang, Xiaotong Jiang, Rong Xiao, and Shoushan Li. 2022. Aspect-based sentiment analysis with opinion tree generation. In *IJCAI*, volume 2022, pages 4044–4050.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. Mvp: Multi-view prompting improves aspect sentiment tuple prediction. *arXiv preprint arXiv:2305.12627*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Mengting Hu, Yinhao Bai, Yike Wu, Zhen Zhang, Liqi Zhang, Hang Gao, Shiwan Zhao, and Minlie Huang. 2023. Uncertainty-aware unlikelihood learning improves generative aspect sentiment quad prediction. *arXiv preprint arXiv:2306.00418*.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. *arXiv preprint arXiv:2210.10291*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in Neural Information Processing Systems*, 35:4328–4343.

Xin Li, Lidong Bing, Piji Li, Wai Lam, and Zhimou Yang. 2018. Aspect term extraction with history attention and selective transformation. *arXiv preprint arXiv:1805.00760*.

Yunlong Liang, Fandong Meng, Jinchao Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2020. An iterative multi-knowledge transfer network for aspect-based sentiment analysis. *arXiv preprint arXiv:2004.01935*.

Jian Liu, Zhiyang Teng, Leyang Cui, Hanmeng Liu, and Yue Zhang. 2021. Solving aspect category sentiment analysis as a text generation task. *arXiv preprint arXiv:2110.07310*.

Pengfei Liu, Shafiq Joty, and Helen Meng. 2015. Fine-grained opinion mining with recurrent neural networks and word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1443.

Yue Mao, Yi Shen, Jingchao Yang, Xiaoying Zhu, and Longjun Cai. 2022. Seq2path: Generating sentiment tuples as paths of a tree. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2215–2225.

Samuel Mensah, Kai Sun, and Nikolaos Aletras. 2021. An empirical study on leveraging position embeddings for target-oriented opinion words extraction. *arXiv preprint arXiv:2109.01238*.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8600–8607.

Joseph J Peper and Lu Wang. 2022. Generative aspect-based sentiment analysis with contrastive learning and expressive structure. *arXiv preprint arXiv:2211.07743*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR.

Stéphan Tulkens and Andreas van Cranenburgh. 2020. Embarrassingly simple unsupervised aspect extraction. *arXiv preprint arXiv:2004.13580*.

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

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9122–9129.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, and Yi Chang. 2021. Eliminating sentiment bias for aspect-level sentiment classification with unsupervised opinion extraction. *arXiv preprint arXiv:2109.02403*.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. *arXiv preprint arXiv:2010.04640*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2018. Double embeddings and cnn-based sequence labeling for aspect extraction. *arXiv preprint arXiv:1805.04601*.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021. Learning span-level interactions for aspect sentiment triplet extraction. *arXiv preprint arXiv:2107.12214*.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. *arXiv preprint arXiv:2010.02609*.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. *arXiv preprint arXiv:2110.00796*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022a. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering (Volume: 35, Issue: 11)*.

Yice Zhang, Yifan Yang, Yihui Li, Bin Liang, Shiwei Chen, Yixue Dang, Min Yang, and Ruifeng Xu. 2022b. Boundary-driven table-filling for aspect sentiment triplet extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6485–6498.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. Spanmlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3239–3248.

Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao. 2015. Representation learning for aspect category detection in online reviews. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Yuxiang Zhou, Lejian Liao, Yang Gao, Zhanming Jie, and Wei Lu. 2021. To be closer: Learning to link up aspects with opinions. *arXiv preprint arXiv:2109.08382*.

Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. 2023. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325.