applied sciences

MDPI

Article

TempCL-Cite: Leveraging Temporal Signals and LLM Positives in Contrastive Learning for Citation Intent Classification

Yu Song, Xiaodong Lu, Lidan Wan * and Zhao Zhang

School of Computer Science and Engineering, Beihang University, Xueyuan Road No.37, Beijing 100191, China; song_yu@buaa.edu.cn (Y.S.); xiaodonglu@buaa.edu.cn (X.L.); zhao_zhang@buaa.edu.cn (Z.Z.)
* Correspondence: zy2306223@buaa.edu.cn

Abstract

Citation intent classification identifies the underlying functions of scholarly citations and provides semantic information beyond citation counts. However, existing methods pay limited attention to temporal information in citation relationships and are often constrained by the scarcity of high-quality annotated data. To address these challenges, we propose TempCL-Cite, a temporal-aware contrastive learning framework that jointly exploits publication time gaps and large language model (LLM)-generated positive examples. TempCL-Cite encodes the publication time gap and fuses it with textual representations, while using temporal-gap differences together with label differences to guide negative sampling. In addition, a prompt-based LLM generates paraphrases of citation contexts conditioned on their intent labels, providing additional positive views for contrastive learning. The LLM is used only for offline data augmentation and is not required during inference. Experiments on three benchmark datasets show that TempCL-Cite achieves the highest mean Accuracy and Macro-F1 among the compared methods. Ablation results indicate that temporal modeling and contrastive learning contribute to performance in the evaluated settings. Furthermore, temporal-gap-constrained negative sampling outperforms both label-only selection and matched-number random sampling on all three datasets. Comparisons with same-label positives and dropout views show that the proposed LLM-based positive construction strategy achieves the highest mean scores among the evaluated alternatives. BERTScore results indicate strong semantic similarity between the generated and original citation contexts.

Keywords: citation intent classification; temporal contrastive learning; large language models; data augmentation; citation networks; deep learning; scholarly data analysis

Check for updates

Academic Editor: Iwona Grobelna

Received: 8 October 2026
Revised: 30 October 2026
Accepted: 30 October 2026
Published: 2 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Citation intent classification, also referred to as citation function classification, aims to automatically identify the underlying purposes or functions of citations in scholarly documents using natural language processing techniques [1]. Rather than treating a citation as a binary link between two publications, citation intent classification seeks to understand why a particular work is cited and what role it plays in the citing document. Such fine-grained semantic information can reveal whether a cited work provides background knowledge, supports a methodological choice, enables comparison, or is directly applied or extended. This perspective is valuable for information retrieval and scholarly recommendation, where distinguishing papers according to their citation functions can facilitate more targeted retrieval and recommendation. For example, researchers may be interested not simply in papers that cite a particular method, but specifically in papers that apply, extend, or compare

that method. Citation intent classification also provides useful information for scientometric analysis. Conventional evaluation indicators, such as citation counts, impact factors, and the h-index, primarily quantify the frequency of scholarly influence but generally do not distinguish the different functions served by individual citations [2]. Since citations can play substantially different rhetorical and intellectual roles, incorporating citation intent can provide a more fine-grained characterization of scholarly relationships and research influence [3].

Early studies on citation intent identification mainly relied on handcrafted linguistic and bibliometric features combined with conventional machine learning algorithms, such as Random Forest [4]. With the development of neural representation learning, deep learning methods based on distributed word representations enabled models to automatically learn task-specific semantic features from citation contexts [5,6]. The emergence of pre-trained language models further improved citation intent classification by providing contextualized representations learned from large-scale corpora. In particular, scientific-domain models such as SciBERT have demonstrated the effectiveness of domain-adapted pre-trained representations for scholarly text classification [7,8]. More recent approaches have explored structural information beyond individual citation contexts. For example, graph neural network methods such as GraphCite incorporate citation-network relationships to model the structural dependencies among scholarly documents [9]. Meanwhile, the rapid development of LLMs has introduced new possibilities for citation intent classification through zero-shot reasoning and parameter-efficient fine-tuning approaches such as LoRA [10–12]. In parallel, contrastive learning has been increasingly investigated as a means of learning more discriminative and generalizable representations by explicitly modeling relationships among semantically related instances [13].

Despite these advances, two challenges remain insufficiently explored in existing citation intent classification methods. First, the temporal information associated with citation relationships has received relatively limited attention. The time gap between a citing paper and a cited paper provides contextual information about the relative temporal position of the two works and may contain useful signals for distinguishing citation functions. Citation relationships involving recently published studies may arise in contexts such as methodological comparison, application, or extension, whereas older publications may more frequently serve as established background or foundational references. However, most existing approaches primarily focus on the textual content of citation contexts, paper-level semantic representations, or citation-network structure, without explicitly incorporating the publication time gap into the learning process. As a result, potentially informative temporal signals remain underutilized.

Second, the limited availability of high-quality annotated data remains a challenge for learning robust citation intent representations. Citation intent annotation requires domain knowledge and careful interpretation of the relationship between a citation context and the cited work, making large-scale manual annotation costly. Consequently, existing benchmark datasets are generally limited in size. Data scarcity is particularly relevant to contrastive learning because the effectiveness of contrastive objectives depends on constructing sufficiently diverse and semantically meaningful positive and negative pairs. Simply duplicating or randomly perturbing existing citation contexts may introduce semantic noise or fail to preserve the original citation intent. Therefore, an effective approach is needed to construct additional training examples while maintaining the semantic and functional consistency of the original citation.

To address these challenges, we propose TempCL-Cite, a temporal-aware contrastive learning framework for citation intent classification that jointly exploits publication time gaps and LLM-generated intent-preserving positive examples. Unlike approaches that

incorporate temporal metadata solely as additional input features, TempCL-Cite uses the relative publication-year gap in two complementary ways. First, it integrates the temporal gap into the representation of a citation context, enabling the classifier to consider temporal information alongside contextual semantics. Second, it uses temporal-gap information in contrastive sample selection, so that negative examples are selected according to both intent-label differences and temporal differences. In addition, LLM-generated paraphrases are introduced as training-time positive examples to enrich the linguistic realizations of citation contexts while retaining their intended citation functions. These components form a task-oriented framework for citation intent classification rather than a simple combination of independent modules. The contribution of TempCL-Cite lies in this joint design of temporal representation, temporally informed contrastive sampling, and intent-preserving positive-example augmentation.The main contributions of this work are summarized as follows:

- We propose TempCL-Cite, a temporal-aware contrastive learning framework for citation intent classification that jointly considers the semantic content of citation contexts and the relative publication-year gap between citing and cited papers.
- We introduce a dual-role temporal modeling strategy that incorporates the relative publication-year gap into both contextual representation learning and contrastive negative-sample selection. This design allows temporal metadata to influence not only the input representation but also the construction of contrastive training signals.
- We incorporate LLM-generated, intent-preserving paraphrases as additional positive examples during training to improve the diversity of citation-context expressions. The generated examples are used only for training and are not required during inference.
- We evaluate the framework on ACL-ARC, SciCite, and ACT2, using baseline comparisons and ablation studies to examine its performance and the contributions of its principal components.

The remainder of this paper is organized as follows. Section 2 reviews related work on citation intent classification, including context representation, auxiliary information, and recent contrastive learning and LLM-based approaches. Section 3 presents the proposed TempCL-Cite framework, covering the problem formulation, the dual-role temporal modeling mechanism, the LLM-based positive-example augmentation strategy, and the training objective. Section 4 reports the experimental setup, main results, ablation studies, and further analyses of negative sampling, positive-pair construction, and semantic fidelity. Section 5 concludes the paper and discusses directions for future work.

2. Related Work

Citation intent classification has progressed substantially over the past two decades, driven by advances in natural language processing (NLP) and the increasing demand for fine-grained analysis of scholarly communication. Existing research can be broadly organized into three interconnected directions: (1) the evolution of citation-context representation from handcrafted features to contextualized language representations; (2) the incorporation of auxiliary information beyond the citation context, including citation-network structure and multi-source supervision; and (3) the recent exploration of contrastive learning and LLMs for improving representation learning and alleviating data limitations.

Context Representation. Early research on citation intent classification primarily relied on rule-based and feature-engineered approaches. The automatic system proposed by Teufel et al. [14] employed an IBK classifier with syntactic and lexical features, including cue phrases such as “we adopt” and “in contrast to”, establishing an early benchmark for automated citation-function classification. Subsequently, Abu-Jbara et al. [15] systematically investigated structural features, including section headings and sentence positions, together with lexical n-grams, and showed that Support Vector Machines (SVMs) could

effectively classify citation functions. Similarly, Jurgens et al. [4] incorporated positional and discourse-related features into an SVM-based framework, demonstrating the difficulty of distinguishing fine-grained citation functions even with rich handcrafted features.

The development of neural representation learning marked an important transition from manually designed features toward automatically learned semantic representations. Hassan et al. [16] applied Long Short-Term Memory (LSTM) networks to citation intent classification, demonstrating the potential of neural sequence models for learning task-specific representations directly from citation contexts. Cohan et al. [17] introduced Structural Scaffolds, a multi-task learning framework that jointly optimized citation intent classification with auxiliary tasks such as section prediction using an attention-based BiLSTM. Similarly, Yousif et al. [18] developed a Recurrent Convolutional Neural Network (RCNN) in a multi-task setting to jointly model citation purpose and sentiment. These studies demonstrated that neural models could capture contextual and semantic patterns that were difficult to represent using conventional feature-engineering approaches.

The emergence of pre-trained language models (PLMs) further improved the representation of scholarly text. General-domain BERT [19] and scientific-domain models such as SciBERT [7] provide contextualized representations that can be adapted to citation intent classification through task-specific fine-tuning. Roman et al. [20] employed BERT to encode citation contexts and reported improvements over earlier neural approaches. Maheshwari et al. [21] further demonstrated that SciBERT representations alone could achieve competitive citation intent classification performance without relying on highly complex task-specific architectures. More recently, domain-adapted and pre-trained language models have continued to serve as strong foundations for citation-related classification tasks.

Auxiliary Information. Although citation-context representation provides important semantic information, citation intent can also be influenced by information beyond the citing sentence. Existing studies have therefore explored auxiliary signals from two major perspectives: citation-network structure and additional supervision from related tasks or datasets.

Structural signals exploit the citation graph, in which scholarly papers are represented as nodes and citation relationships as edges. Rather than treating each citation independently, graph-based approaches model relational information among citing and cited papers. Berrebbi et al. [9] proposed GraphCite, which combines citation-network embeddings obtained through Graph Convolutional Networks (GCNs) with PLM-based textual representations. Their results demonstrated that citation-network structure can provide complementary information when textual evidence alone is insufficient. Tsai et al. [22] constructed a more comprehensive citation graph containing supporting evidence spans and employed a Gated Graph Convolution Network to model relationships among citation-related entities. Du et al. [23] further explored a text-free graph-based approach that learns citation intent from weakly labeled structural information, highlighting the potential of network-level signals in low-resource settings.

Beyond structural information, researchers have also investigated additional supervision from related tasks, datasets, and domains. Multi-task learning (MTL) provides one representative approach. As discussed above, Cohan et al. [17] incorporated auxiliary tasks such as section prediction to regularize citation intent classification. Qi et al. [24] developed a BiLSTM-SciBERT multi-task framework that jointly learns citation intent and related metadata prediction, improving performance across citation-intent categories. Multi-dataset training has also been investigated as a means of learning more generalizable representations. Shui et al. [25] fine-tuned PLMs on multiple citation intent datasets and found that joint training can reduce dataset-specific biases and improve cross-dataset generalization. However, citation-intent datasets differ considerably in their annotation schemes, ranging

from relatively coarse categories to substantially more fine-grained taxonomies. As noted by Zhang et al. [26], the absence of a unified annotation standard remains an important challenge for fair comparison and broad model transfer.

Contrastive Learning and LLMs. In addition to improving model architecture and incorporating auxiliary information, recent research has increasingly focused on data-efficient representation learning. Citation intent datasets are often limited in size, while the distribution of intent categories can also be highly imbalanced. These characteristics make it challenging to learn robust representations, particularly for fine-grained and relatively infrequent intent categories. Contrastive learning provides a natural framework for addressing this challenge by explicitly encouraging semantically related instances to have similar representations while separating instances associated with different labels.

Recent studies have applied contrastive learning to citation-related tasks by constructing positive and negative pairs from existing training instances. For example, Li et al. [13] explored contrastive learning for citation intent classification and demonstrated its potential for improving representation quality under limited supervision. However, constructing reliable positive examples remains non-trivial. Conventional augmentation techniques, such as sentence cropping, synonym replacement, or keyword perturbation, may alter citation-specific semantic cues. In citation intent classification, seemingly minor changes to a citation context can modify the rhetorical function being expressed. For example, removing an intent-bearing phrase may change the distinction between a methodological citation and a comparative citation. Therefore, augmentation strategies for citation intent classification need to balance textual diversity with preservation of the original citation intent.

The rapid development of LLMs provides new opportunities for both citation intent classification and data augmentation. Existing studies have primarily investigated LLMs in two ways: directly performing citation intent classification through prompting or fine-tuning, and generating additional training data. Kunnath et al. [27] showed that few-shot prompting with GPT-3 can achieve competitive citation intent classification performance without parameter updates, although performance can be sensitive to prompt design and task formulation. Karaca and Eravcı [11] investigated the fine-tuning of open-source LLMs, such as LLaMA, for citation intent classification and reported strong classification performance, while also highlighting the associated computational requirements. Koloveas et al. [12] systematically evaluated multiple LLMs and found that their performance can vary across citation-intent taxonomies, with relatively fine-grained and less frequent categories remaining challenging. Fogelson et al. [10] further discussed limitations of LLM-based citation analysis, including the reliability of generated reasoning. These studies demonstrate the potential of LLMs for citation intent classification, while also suggesting that their use as auxiliary data-generation tools may provide an alternative to relying on LLM inference at prediction time.

Research Gap and Positioning of Our Work. The above studies have substantially advanced citation intent classification, yet several aspects remain insufficiently explored. First, although citation relationships inherently involve a temporal dimension, publication time gaps have received comparatively limited attention in existing citation intent classification frameworks. Most approaches focus primarily on textual representations, network structure, or auxiliary supervision, leaving the temporal distance between citing and cited publications underutilized. Second, existing contrastive learning approaches require reliable positive and negative pairs, while conventional text augmentation may fail to preserve citation intent when modifying fine-grained scientific language. LLMs provide a promising mechanism for generating semantically consistent textual variants, but their potential as intent-preserving positive-example generators within a citation-intent contrastive learning framework remains underexplored.

To address these gaps, we propose TempCL-Cite, which jointly incorporates temporal information and LLM-generated positive examples into a unified contrastive learning framework. The publication time gap is assigned a dual role: it guides the construction of informative negative pairs and is explicitly fused with textual representations to enrich citation representations. Meanwhile, a prompt-based LLM augmentation strategy generates intent-preserving positive examples that are incorporated directly into contrastive learning. Unlike approaches that require LLM inference for each prediction, the LLM in TempCL-Cite is used during data augmentation, while the final classifier operates without LLM-based inference. By integrating these two complementary sources of supervision within a single learning framework, TempCL-Cite aims to improve citation intent representations under limited annotated data.

3. Methodology

3.1. Problem Formulation

Given a scientific publication, we focus on classifying the intent of each individual citation within it. For a target citation i, its citation context T_i is defined as a contiguous span of text that contains the citation anchor (e.g., "[1]") in its original form. The task of citation intent classification aims to predict the intent label $y_i \in C$ for the context T_i, where C is a predefined set of intent categories (e.g., *Background*, *Comparison*). In addition to T_i, our model leverages the citation time gap Δt_i between the citing and cited papers, defined as the publication-year difference:

$$\Delta t_i = t_i^{\text{cite}} - t_i^{\text{cited}},$$

where t_i^{cite} and t_i^{cited} denote the publication years of the citing and cited papers, respectively. Unlike absolute publication years, Δt_i captures the relative temporal distance between the citing and cited papers. We also use an LLM-generated positive example derived from T_i.

3.2. Model Architecture

The overall architecture of TempCL-Cite is illustrated in Figure 1. The framework is designed around four key elements: (1) the original citation context, (2) the corresponding publication time gap Δt, (3) an LLM-augmented positive example, and (4) in-batch negative examples. It learns through two synergistic branches: a contrastive learning branch for representation learning and a classification branch for final intent prediction. The two branches share the same SciBERT encoder.

3.2.1. Dual-Role Temporal Utilization Mechanism

The temporal signal Δt (the publication time gap) is utilized in two distinct yet complementary ways within TempCL-Cite.

Feature Fusion for Classification. The scalar time gap $\Delta t \in \mathbb{R}$ is first projected into a dense feature vector to enrich the semantic representation. Specifically, it is transformed via a learnable feed-forward layer:

$$\mathbf{f}_t = \text{GELU}(\mathbf{W}_t \Delta t + \mathbf{b}_t), \tag{1}$$

where $\mathbf{W}_t \in \mathbb{R}^{d_t \times 1}$ and $\mathbf{b}_t \in \mathbb{R}^{d_t}$ are trainable parameters, and d_t denotes the temporal embedding dimension. This temporal representation $\mathbf{f}_t$ is then concatenated with the semantic representation of the original context $\mathbf{Z}_a$—obtained from the [CLS] token embedding of SciBERT—to form the enhanced feature vector for the final classifier:

$$\mathbf{h}_{\text{fused}} = [\mathbf{Z}_a; \mathbf{f}_t]. \tag{2}$$

Temporal-Guided Negative Sampling for Contrastive Learning. To construct a more challenging contrastive task, the temporal signal Δt actively guides the selection of negative samples. In practice, for an anchor sample with label y_a and time gap Δt_a, we define the candidate negative set $\mathcal{N}$ from within the same training batch by applying a dual-criteria filtering strategy. Specifically, a sample j is considered a valid negative only if it satisfies both:

$$y_j \neq y_a \quad \text{and} \quad \Delta t_j \neq \Delta t_a, \tag{3}$$

while also excluding the anchor itself and its LLM-generated positive counterpart. This enables the negative sampling process to incorporate both semantic and temporal information.

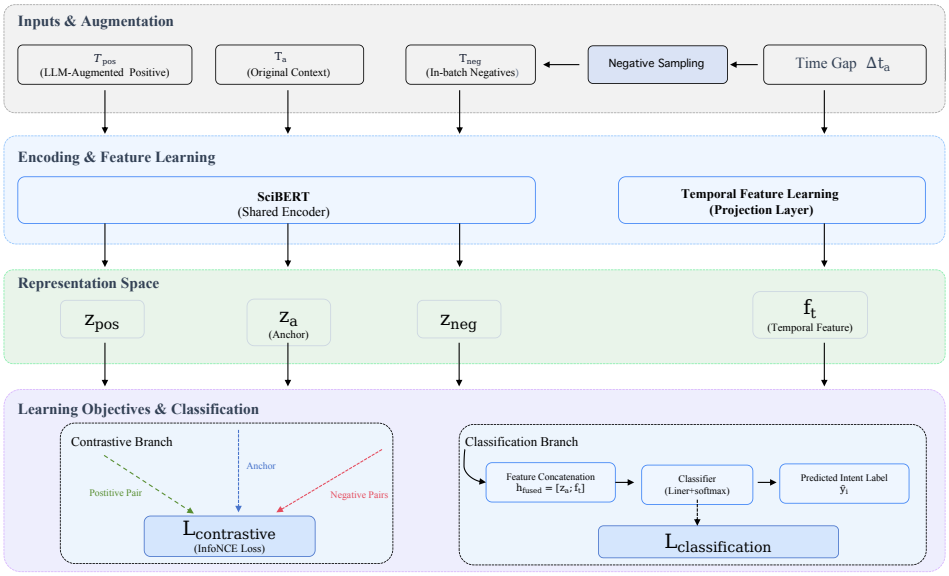


Figure 1. The TempCL-Cite framework. A prompt-guided LLM generates positive examples, while the time gap (Δt) plays a dual role as a fused feature and a guide for negative sampling.

Formally, the negative mask $\mathcal{M}_{\text{neg}}$ for anchor a is defined as:

$$\mathcal{M}_{\text{neg}} = \{j \mid y_j \neq y_a \wedge \Delta t_j \neq \Delta t_a \wedge j \notin \{a, a^+\}\}, \tag{4}$$

where a^+ denotes the positive pair of a. During contrastive learning, only the samples within $\mathcal{M}_{\text{neg}}$ are treated as valid negatives for the anchor. This focuses the learning objective on contrasting against these curated, informative negatives.

3.2.2. Prompt-Based LLM Augmentation for Positive Examples

To generate high-quality positive pairs for contrastive learning without altering the citation intent, we employ a prompt-based paradigm with LLMs. We design a structured prompt template to instruct the LLM (e.g., Qwen 3 [28]) to paraphrase the original citation context while strictly preserving its core factual content and rhetorical function (intent). This approach ensures that the generated text $\mathbf{T}_{pos}$ serves as a semantically consistent and intent-preserving positive example for the anchor context $\mathbf{T}_a$. Compared to [13], our prompt explicitly incorporates the citation intent, aiming to generate more intent-consistent positive examples. The exact prompt template used is shown in Figure 2. The template explicitly conditions the LLM on the original citation text and its corresponding intent label, enforcing both semantic preservation and intent-awareness. The JSON output constraint ensures the generated positive examples can be directly parsed and integrated into the contrastive learning pipeline without post-processing.

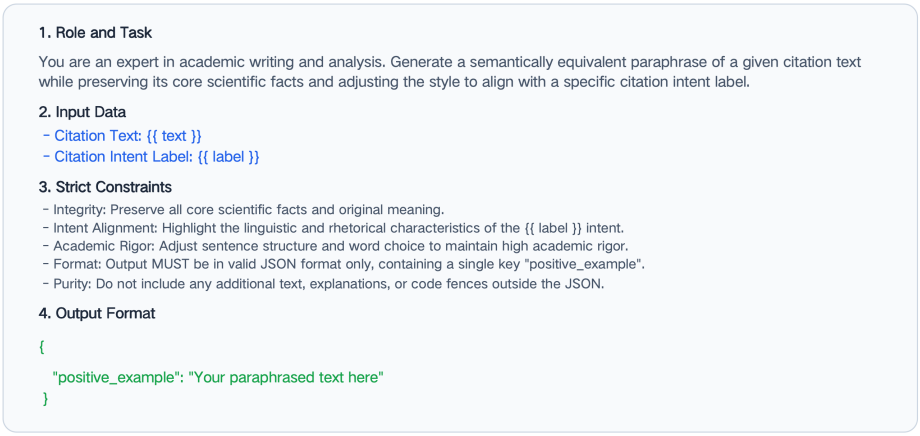


Figure 2. The structured prompt template used for citation paraphrasing.

During generation, we set the temperature to 0.3, a relatively low value that prioritizes semantic fidelity to the original intent while still introducing sufficient lexical and syntactic variation to serve as effective positive pairs. To introduce the stochasticity necessary for generating diverse paraphrases, we enable sampling-based decoding via `do_sample=True`. Greedy decoding is deliberately avoided, as it would produce deterministic and near-identical outputs that offer limited benefit for contrastive learning. The maximum number of newly generated tokens is set to 500, which provides ample capacity to paraphrase even lengthy citation contexts without truncation, ensuring that no critical information is lost. These three parameter choices collectively ensure that the LLM-generated positive examples are both semantically faithful to the original citation intent and sufficiently varied to enrich the contrastive learning signal.

Our augmentation framework is designed with three key principles in mind. First, we strictly prevent label leakage: the ground-truth citation intent labels are used exclusively during training to guide the LLM in generating intent-consistent positive pairs, analogous to the label-based positive selection mechanism in supervised contrastive learning [29]. The LLM is entirely excluded from the inference phase, and all generated samples are strictly confined to the training set, never exposed to the test set during model evaluation. This ensures the rigor of our experimental setup and the trustworthiness of the results. Second, the LLM augmentation is performed entirely offline as a one-time positive sample construction step, incurring only a fixed upfront computational cost. The generated data can be cached and reused across multiple training runs without repeated API calls, making the framework highly efficient for iterative experimentation and hyperparameter tuning. Finally, our framework is agnostic to the underlying LLM implementation: when external API services are unavailable or restricted, it can seamlessly transition to open-source models (e.g., Qwen 3, LLaMA) deployed on local infrastructure, ensuring broad applicability and reproducibility across diverse research environments.

3.2.3. Training Objective

The overall training objective is a weighted sum of the contrastive loss and the classification loss:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{contrastive}} + \beta \cdot \mathcal{L}_{\text{classification}}, \tag{5}$$

where α and β are balancing coefficients.

Contrastive Loss. We apply the InfoNCE loss, pulling the anchor $\mathbf{z}_a$ towards its positive pair $\mathbf{z}_{pos}$ and pushing it away from the temporally-guided negative set $\mathcal{M}_{\text{neg}}(a)$:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{e^{\text{sim}(\mathbf{z}_a, \mathbf{z}_{pos})/\tau}}{e^{\text{sim}(\mathbf{z}_a, \mathbf{z}_{pos})/\tau} + \sum_{j \in \mathcal{M}_{\text{neg}}(a)} e^{\text{sim}(\mathbf{z}_a, \mathbf{z}_j)/\tau}}, \tag{6}$$

where $\mathbf{z}$ denotes the normalized contrastive representation, $\text{sim}(\cdot)$ is cosine similarity, and τ is a temperature hyperparameter.

Classification Loss. The enhanced representation $\mathbf{h}_{\text{fused}}$ is passed through a linear classifier to obtain the classification logits $\boldsymbol{\ell} \in \mathbb{R}^{|C|}$:

$$\boldsymbol{\ell} = \mathbf{W}_c \mathbf{h}_{\text{fused}} + \mathbf{b}_c, \tag{7}$$

where $\mathbf{W}_c \in \mathbb{R}^{|C| \times (d+d_t)}$ and $\mathbf{b}_c \in \mathbb{R}^{|C|}$ denote the weight matrix and bias vector of the classification layer, respectively. Here, d is the dimensionality of the semantic representation $\mathbf{Z}_a$, and d_t is the dimensionality of the temporal representation $\mathbf{f}_t$. The logits are then transformed by the softmax function to obtain the predicted probability distribution $\hat{\mathbf{y}}$ over the intent categories C:

$$\hat{\mathbf{y}} = \text{softmax}(\boldsymbol{\ell}). \tag{8}$$

The classification loss is defined as the cross-entropy between the predicted distribution and the ground-truth one-hot label $\mathbf{y}$:

$$\mathcal{L}_{\text{classification}} = -\sum_{c=1}^{|C|} y_c \log(\hat{y}_c), \tag{9}$$

where $y_c \in \{0, 1\}$ denotes the c-th component of the one-hot ground-truth label $\mathbf{y}$, and $\hat{y}_c$ denotes the predicted probability assigned to class c.

3.3. Use of Generative Artificial Intelligence

ChatGPT (OpenAI, San Francisco, CA, USA) was used during manuscript preparation to assist with academic writing, including English grammar correction, sentence structure refinement, and improvements to clarity and readability. In addition, the Qwen3-series model Qwen-Plus (version: `qwen-plus-2025-07-14`, provided by Alibaba Cloud Model Studio) was employed as part of the proposed TempCL-Cite methodology to generate intent-preserving paraphrases of citation contexts for contrastive learning.

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

Our evaluation covers three widely used citation intent classification benchmarks: ACL-ARC [4], SciCite [17], and ACT2 [30]. Specifically, ACL-ARC contains six citation intent classes: *Background*, *CompareOrContrast*, *Extends*, *Future*, *Motivation*, and *Uses*. SciCite contains three classes: *Background*, *Method*, and *Result*. ACT2 contains six citation intent classes: *Background*, *CompareOrContrast*, *Extends*, *Future*, *Motivation*, and *Uses*. Table 1 summarizes the number of citations and the size of the intent label space for each benchmark. To facilitate temporal analysis, we augment the datasets with paper publication years. For SciCite, we obtain the year of each publication via the Semantic Scholar API. For ACT2, publication years are retrieved using the CrossRef API. The ACL-ARC dataset includes year information natively.

Table 1. Overview of the three evaluation datasets.

Dataset	Citations	Intent Classes
ACL-ARC	1941	6
SciCite	11,020	3
ACT2	4000	6

4.1.2. Temporal Metadata Collection and Processing

Temporal metadata were obtained using dataset-provided fields or external bibliographic APIs, depending on the dataset. For ACL-ARC, the publication years of the citing and cited papers were directly obtained from the existing `citing_paper_year` and `cited_paper_year` fields. For SciCite, the `citingPaperId` and `citedPaperId` fields were used to retrieve the publication years of the citing and cited papers, respectively, through the Semantic Scholar Graph API. For ACT2, the Crossref API was queried using the titles of the citing and cited papers. The `published` field was prioritized, and `published-online` was used as a fallback when the year field was unavailable. Temporal metadata were successfully obtained for all samples across the three datasets, ensuring that the publication-year information required for temporal modeling was available for every sample.

4.1.3. Evaluation Metrics

We report Accuracy and Macro-averaged F1 to assess citation intent classification performance. Accuracy measures the overall proportion of correctly predicted instances. Macro-F1 first computes the F1 score for each intent category independently and then averages these scores, so every category contributes equally; this makes it more robust when the class distribution is imbalanced.

4.1.4. Baseline Methods

We compare our proposed method against the following representative and state-of-the-art baselines:

- LLM Zero-Shot: We utilize a large language model in a zero-shot prompting setting without any task-specific training. Specifically, we use the Qwen-Turbo model (via the Alibaba Cloud API) to directly classify the citation intent. The prompt takes the citation text and the available intent categories as input, and asks the model to output the citation intent label. The candidate categories are specified according to the label set of the dataset. The predicted label is the intent generated by the model. This baseline evaluates the ability of a general-purpose LLM to perform citation intent classification without fine-tuning. The prompt template used for this baseline is illustrated in Figure 3.
- LLM LoRA Fine-tuning: We select the compact Qwen/Qwen3-0.6B model for computational efficiency, and fine-tune it using Low-Rank Adaptation (LoRA) [31] with rank $r = 8$, scaling factor $\alpha = 16$, dropout 0.05, and adapters applied to the query, key, value, and output projection matrices. We use a single-label classification prompt that contains an instruction, the available categories, the citation text, and a `Category:` field; the model is trained to generate the gold label. This represents a parameter-efficient approach to adapting LLMs for this specific task. The prompt template is illustrated in Figure 4.

```
1. Role and Task

You are an expert in scientific literature analysis. Identify the citation intent of a given citation text based on its semantic meaning and function in the scientific discourse.

2. Input Data

- Candidate Intent Categories: {{ categories }}
- Citation Text: {{ text }}

3. Strict Constraints

-Classification: Select exactly one intent label from the provided candidate categories.
-Semantic Accuracy: Determine the intent based on the meaning and function of the citation text.
-Label Consistency: Use a label exactly as specified in the candidate categories.
-Output: Return only the predicted intent label without additional explanations.

4. Output Format

{
  "intent": "predicted_label"
}
```

Figure 3. Prompt template used for the LLM Zero-Shot baseline.

```
1. Instruction

You are an expert in academic literature analysis. Please analyze the citation text below and determine the author's citation intent.

2. Available Categories

{{ labels}}

3. Citation Text

{{ text}}

4. Output Prefixt

Category:
```

Figure 4. Prompt template used for the LLM LoRA Fine-Tuning baseline.

- SciBERT: We include a fine-tuned SciBERT [7] as a strong domain-specific pre-trained language model baseline for citation intent classification.
- GraphCite: We adopt GraphCite [9], a graph neural network model designed to exploit citation graph structure for intent classification. Experiments are conducted using its official implementation released in the GitHub repository (https://github.com/nyxpho/graphcite, accessed on 29 October 2026).
- CiteFusion: We include CiteFusion [8], an ensemble framework that transforms multi-class citation intent classification into multiple binary classification tasks using a one-vs-all strategy. For each intent class, SciBERT [7] and XLNet [32] are independently fine-tuned as complementary classifiers, and their predictions are subsequently aggregated to determine the final citation intent label.
- Citss: We include Citss [13] as a baseline, which uses self-supervised contrastive learning for citation classification. We use the official implementation available on GitHub (https://github.com/LITONG99/Citss, accessed on 29 October 2026).

4.1.5. Implementation Details.

We implement TempCL-Cite using the Transformers library, with SciBERT as the backbone for encoding citation contexts. The key parameter settings are as follows: the temporal feature projection dimension is 32, the contrastive learning projection dimension is 256, the loss weighting factors are set to $\alpha = 0.3$ and $\beta = 1$ (for the contrastive and

classification losses, respectively), the maximum sequence length is 256 tokens, the batch size is 64, the learning rate is 2×10^{-5}, the early stopping patience is 5 epochs, and the dropout rate across the entire network is 0.2. This configuration is designed to effectively evaluate the benefits of joint temporal and contrastive learning.

To account for stochastic variability in deep-learning experiments, all methods are evaluated over three independent runs with different random seeds, and the reported results are averaged across these runs. Previous studies have highlighted the importance of accounting for experimental variability and ensuring reproducible evaluation in machine learning [33]. Related empirical studies have also examined model performance under different operational conditions and data configurations, highlighting the importance of evaluating models in their respective application contexts [34,35]. Accordingly, the reported mean scores are interpreted as empirical comparisons under the evaluated settings, rather than as evidence of statistically significant differences.

4.2. Results and Analysis

4.2.1. Main Results

We report the overall performance of all compared methods on the three benchmark datasets in Table 2. All scores are averaged over three independent runs with different random seeds. As shown, TempCL-Cite achieves the highest mean scores across all datasets and evaluation metrics among the compared methods, providing empirical support for the effectiveness of the proposed framework under the evaluated settings. The largest numerical gains are observed on the ACL-ARC dataset (e.g., +5.75 percentage points in accuracy and +4.13 percentage points in Macro-F1 over the strongest baseline). Steady gains are also observed on the SciCite and the cross-domain ACT2 datasets, confirming the effectiveness and generalization ability of our approach.

Table 2. Overall performance comparison (%). Best results in **bold**.

Model	ACL-ARC		SciCite		ACT2	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
LLM Zero-Shot	53.96	25.03	68.57	62.88	59.50	14.59
LLM LoRA	75.54	63.71	86.24	84.48	62.33	38.09
SciBERT	76.26	67.54	85.71	84.16	61.67	39.17
GraphCite	74.91	72.99	86.24	86.24	63.18	41.96
CiteFusion	76.26	68.42	88.17	86.56	63.50	41.35
Citss	74.86	67.53	86.32	85.10	63.58	41.92
TempCL-Cite	**82.01**	**77.12**	**88.18**	**86.92**	**64.33**	**42.83**

A closer examination of the results reveals that the magnitude of the performance improvement varies across datasets. The most substantial gains are observed on ACL-ARC. Specifically, TempCL-Cite outperforms the strongest baseline by 5.75 percentage points in accuracy (82.01% vs. 76.26%, SciBERT) and 4.13 percentage points in Macro-F1 (77.12% vs. 72.99%, GraphCite). These results show that the combination of temporal contrastive learning and LLM-based positive augmentation yields clear improvements on this benchmark.

On SciCite, TempCL-Cite obtains an accuracy of 88.18% and a Macro-F1 of 86.92%. These are the highest mean scores among the evaluated methods. The accuracy difference between TempCL-Cite and CiteFusion is only 0.01 percentage points (88.18% vs. 88.17%), whereas the Macro-F1 difference is 0.36 percentage points (86.92% vs. 86.56%). In particular, the small accuracy difference should not be interpreted as evidence of a meaningful advantage without further statistical analysis.

The ACT2 dataset also shows consistent improvements. TempCL-Cite achieves 64.33% accuracy, improving over the strongest baseline in accuracy (Citss: 63.58%) by 0.75 percentage points, while its Macro-F1 of 42.83% exceeds the strongest baseline in this metric (GraphCite: 41.96%) by 0.87 percentage points. Although the gains are smaller than those observed on ACL-ARC, the improvements are consistent across both metrics, indicating that the proposed framework remains effective under different dataset sizes and domain characteristics.

Overall, these results show that TempCL-Cite achieves the best mean performance across all three benchmark datasets and metrics, while maintaining consistent improvements over the strongest baselines. The magnitude of the improvement varies across datasets; however, given the limited number of datasets considered, we do not draw conclusions about the factors underlying this variation. Instead, the results provide empirical evidence that temporal-aware contrastive learning and LLM-based positive augmentation can complement conventional supervised learning and improve citation intent classification performance across the evaluated benchmarks.

4.2.2. Ablation Study

To dissect the contribution of each key component in our framework, we conduct ablation experiments on all three datasets. As shown in Table 3, we compare the full TempCL-Cite model against three variants: (1) removing the temporal projector module (w/o temporal projector), (2) removing the contrastive learning head (w/o contrastive head), and (3) removing both components (w/o both additions), which essentially reduces to the SciBERT baseline fine-tuned solely with the classification loss.

Table 3. Ablation study (Accuracy and Macro-F1, %) across ACL-ARC, SciCite, and ACT2 datasets. Best results for each metric and dataset are in **bold**.

Model	ACL-ARC		SciCite		ACT2	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Full TempCL-Cite	**82.01**	**77.12**	**88.18**	**86.92**	**64.33**	**42.83**
w/o temporal projector	81.29	74.37	87.32	86.10	62.33	42.35
w/o contrastive head	80.50	73.65	87.43	86.21	63.00	42.73
w/o both additions	76.26	67.54	85.71	84.16	61.67	39.17

The results clearly demonstrate that both proposed components contribute positively to the final performance. On ACL-ARC, removing the temporal projector leads to a drop of 0.72% in accuracy and 2.75% in Macro-F1, indicating that modeling time gaps between citing and cited papers provides valuable discriminative signals, especially for fine-grained intent distinctions. Removing the contrastive head causes a reduction of 1.51% in accuracy and 3.47% in Macro-F1, highlighting the importance of the contrastive learning objective in shaping a well-structured representation space. When both components are removed, performance degrades substantially to the SciBERT baseline level (76.26% accuracy, 67.54% Macro-F1), confirming that the gains of the full model stem from the synergistic integration of temporal information and contrastive learning rather than from other confounding factors.

A similar pattern holds on the larger SciCite dataset, although the absolute gaps are slightly smaller. The full model achieves 88.18% accuracy and 86.92% Macro-F1. Removing the temporal projector reduces accuracy by 0.86% and Macro-F1 by 0.82%; removing the contrastive head reduces accuracy by 0.75% and Macro-F1 by 0.71%. The complete removal of both components results in a drop of 2.47% in accuracy and 2.76% in Macro-F1 compared to the full model. These consistent degradations indicate that both modules are effective even when sufficient labeled data is available.

On the cross-domain ACT2 dataset, the results show a different pattern from ACL-ARC and SciCite. Removing the temporal projector reduces Accuracy from 64.33% to 62.33% and Macro-F1 from 42.83% to 42.35%, corresponding to decreases of 2.00 and 0.48 percentage points, respectively. In contrast, removing the contrastive head reduces Accuracy to 63.00% and Macro-F1 to 42.73%, corresponding to decreases of 1.33 and 0.10 percentage points. Therefore, under the tested configurations, the temporal projector has a larger measured contribution than the contrastive head on ACT2. This may be because ACT2 is a cross-domain dataset with different citation patterns, where temporal cues are particularly informative. The contrastive head still provides a positive contribution, and the full model achieves the best overall performance, indicating that the two components remain complementary. However, the relative contribution of the contrastive objective is smaller on ACT2 than on ACL-ARC and SciCite.

In summary, the ablation study shows that both the temporal projector and the contrastive learning head contribute positively to TempCL-Cite, although their measured contributions vary across datasets. Removing both components leads to a more substantial performance decline than removing either component individually on all three datasets. These results support the effectiveness of integrating temporal information and contrastive learning within a unified framework.

4.2.3. Analysis of Temporal-Guided Negative Sampling

The temporal gap Δt enters TempCL-Cite in two ways: through the temporal feature projector and through the construction of contrastive negatives. To test whether the latter use matters, we compare three sampling strategies: *Label-only*, *Label + Random Sampling (Matched Number)*, and *Label + Temporal Gap*. The comparison is designed to separate the effect of temporal guidance from the effect of changing the number of negatives.

In the first strategy, denoted as *Label-only*, a sample j is selected as a negative for anchor a if its citation intent differs from that of the anchor:

$$\mathcal{N}_{\text{label}}(a) = \left\{ j \mid y_j \neq y_a,\ j \notin \{a, a^+\} \right\}. \tag{10}$$

This strategy ignores temporal gaps, so candidates satisfying $y_j \neq y_a$ but $\Delta t_j = \Delta t_a$ remain eligible as negatives.

In the second strategy, denoted as *Label + Temporal Gap*, we additionally require the temporal gaps of the anchor and candidate negative to be different:

$$\mathcal{N}_{\text{temporal}}(a) = \left\{ j \mid y_j \neq y_a \wedge \Delta t_j \neq \Delta t_a,\ j \notin \{a, a^+\} \right\}. \tag{11}$$

This constraint removes candidates whose relative publication-year gap matches that of the anchor. A different temporal gap is not assumed to make a negative sample intrinsically harder or more informative; the constraint mainly changes which temporal relations appear in the selected negative pairs. It also shrinks the candidate pool. Therefore, a comparison between *Label-only* and *Label + Temporal Gap* alone confounds temporal selection with the number of negative samples.

To control for the number of negatives, the third strategy, *Label + Random Sampling (Matched Number)*, first constructs the label-constrained candidate set $\mathcal{N}_{\text{label}}(a)$ and then randomly selects a subset without replacement. The subset size is matched to that of the temporal-guided strategy:

$$\begin{aligned} &\mathcal{N}_{\text{random}}(a) \subseteq \mathcal{N}_{\text{label}}(a), \\ &|\mathcal{N}_{\text{random}}(a)| = \left|\mathcal{N}_{\text{temporal}}(a)\right|. \end{aligned} \tag{12}$$

The random baseline imposes no temporal-gap constraint, but it uses the same number of negatives per anchor as the temporal-guided strategy. The two matched strategies therefore differ in their selection criterion rather than in the negative count.

All three strategies share the same SciBERT encoder, temporal feature projector, LLM-generated positive examples, contrastive projection head, classification branch, data splits, and optimization settings. Only the negative sampling strategy differs. For the random baseline, selection follows the same sampling schedule and per-anchor negative count as the temporal-guided strategy.

Table 4 shows that *Label + Temporal Gap* achieves the best Accuracy and Macro-F1 on all three datasets. Against *Label-only*, the gains are 0.72 and 2.79 percentage points on ACL-ARC, 1.77 and 1.62 points on SciCite, and 0.33 and 1.43 points on ACT2. The direction is consistent across datasets, although the Accuracy gain on ACT2 is small.

Table 4. Comparison of different negative sampling strategies (Accuracy and Macro-F1, %). Best results for each dataset and metric are in **bold**.

Negative Sampling	ACL-ARC		SciCite		ACT2	
	Accuracy	Macro-F1	Accuracy	Macro-F1	Accuracy	Macro-F1
Label-only	81.29	74.33	86.41	85.30	64.00	41.40
Label + Random Sampling (Matched Number)	79.14	72.25	86.30	85.22	61.17	41.07
Label + Temporal Gap	**82.01**	**77.12**	**88.18**	**86.92**	**64.33**	**42.83**

The matched-number comparison is more informative. *Label + Temporal Gap* outperforms *Label + Random Sampling (Matched Number)* on all six dataset–metric combinations. On ACL-ARC, the temporal-guided strategy gains 2.87 points in Accuracy and 4.87 points in Macro-F1 over random sampling. The corresponding gains are 1.88 and 1.70 points on SciCite, and 3.16 and 1.76 points on ACT2. Since the negative count per anchor is held fixed, these gaps cannot be attributed to using fewer negatives. They point to the selection criterion itself. At the same time, random sampling is slightly below *Label-only* on ACL-ARC and SciCite, and clearly below it on ACT2 Accuracy, so discarding eligible negatives at random does not help.

One plausible reading is that the temporal constraint changes the temporal relations represented in negative pairs: candidates with the same relative gap as the anchor are excluded even when their labels differ. Together with the temporal feature projector, this may expose the contrastive objective to a wider range of temporal configurations. The aggregate metrics cannot identify the exact mechanism, but they are consistent with this interpretation.

Taken together, the comparison between *Label-only* and *Label + Temporal Gap* estimates the overall effect of the temporal constraint, while the matched-number random baseline holds the negative count fixed. The advantage over both alternatives supports using temporal-gap information in contrastive negative sampling for TempCL-Cite, without assuming that different temporal gaps are intrinsically harder negatives.

4.2.4. Analysis of Positive Pair Construction Strategies

We compare TempCL-Cite with two simpler positive-pair construction strategies, retrained under otherwise identical settings. In *Same-Label Positives*, for each training instance, we randomly select another instance from the training set that shares the same citation-intent label but has different text. The selected instance's original text is used verbatim as the positive example, without LLM-based rewriting. Importantly, positive examples are selected exclusively from the training set, and no validation or test instances are used for positive-pair construction. In *Dropout Views*, the positive example is the input itself, and the two views of a positive pair are obtained through two independent forward passes with dropout enabled, following a dropout-based consistency strategy.

Table 5 reports Accuracy and Macro-F1 for the three strategies on ACL-ARC, SciCite, and ACT2. TempCL-Cite with LLM-generated positives achieves the highest scores on both metrics across all three datasets. Compared with Same-Label Positives, it improves Accuracy by 5.83, 0.92, and 1.43 percentage points on ACT2, SciCite, and ACL-ARC, respectively, and improves Macro-F1 by 1.57, 0.56, and 2.07 percentage points. Compared with Dropout Views, the corresponding Accuracy gains are 6.00, 1.40, and 1.51 percentage points, while the Macro-F1 gains are 0.06, 1.17, and 5.67 percentage points. Although Dropout Views achieves a Macro-F1 close to that of TempCL-Cite on ACT2, larger differences are observed on SciCite and, particularly, ACL-ARC.

Table 5. Comparison of positive-pair construction strategies on ACL-ARC, SciCite, and ACT2. The best result for each dataset and metric is shown in bold.

Dataset	Positive-Pair Strategy	Accuracy (%)	Macro-F1 (%)
ACT2	LLM-generated positives (TempCL-Cite)	**64.33**	**42.83**
	Same-Label Positives	58.50	41.26
	Dropout Views	58.33	42.77
SciCite	LLM-generated positives (TempCL-Cite)	**88.18**	**86.92**
	Same-Label Positives	87.26	86.36
	Dropout Views	86.78	85.75
ACL-ARC	LLM-generated positives (TempCL-Cite)	**82.01**	**77.12**
	Same-Label Positives	80.58	75.05
	Dropout Views	80.50	71.45

The two alternative strategies provide different forms of positive pairs. Same-Label Positives exploits label agreement between existing instances but does not explicitly generate new textual variants. Dropout Views, in contrast, introduces stochastic variation at the representation level without adding new textual content. LLM-generated positives provide additional phrasings intended to preserve the original citation intent, potentially exposing the encoder to a broader range of linguistic expressions associated with the same citation function. The consistently higher scores across the three datasets suggest that LLM-generated positives provide a more effective alternative within our experimental setting. However, the magnitude of the improvement varies across datasets, and the observed performance differences alone do not establish the precise mechanism underlying these gains.

4.2.5. Semantic Fidelity Assessment of LLM-Augmented Positive Samples

To assess the semantic fidelity of the LLM-generated positive samples, we employ BERTScore [36] to measure the semantic similarity between the original citation contexts and their corresponding generated paraphrases. The evaluation is conducted on the augmented training sets of the three benchmark datasets. Since the generated samples are introduced as positive instances in the contrastive learning framework, this evaluation aims to determine whether they maintain strong semantic correspondence with the original citation contexts while allowing appropriate linguistic variation.

We employ BERTScore [36] to quantify the semantic similarity between each original citation context and its corresponding LLM-generated paraphrase. BERTScore evaluates semantic correspondence based on contextualized representations derived from pretrained language models. Compared with conventional lexical-overlap metrics, it is more suitable for paraphrased text because it can capture semantic relationships even when the surface forms of the two texts differ.

For each generated positive sample, we independently compute its BERTScore Precision, Recall, and F1 against the corresponding original citation context. We then calculate the mean value of each metric over all generated positive samples within the corresponding dataset. Thus, the reported values reflect the average semantic similarity of the generated samples at the dataset level.

As shown in Table 6, the generated positive samples achieve consistently high BERTScore scores across all three benchmark datasets. The average F1 scores are 0.9300 for ACL-ARC, 0.9431 for SciCite, and 0.9447 for ACT2, with an overall average F1 of 0.9393. The relatively narrow range of F1 scores (0.9300–0.9447) indicates that the semantic fidelity of the generated samples is stable across datasets with different citation-intent categories and characteristics.

Table 6. Average BERTScore over all generated positive samples.

Dataset	Avg. Precision	Avg. Recall	Avg. F1
ACL-ARC	0.9331	0.9284	0.9300
SciCite	0.9390	0.9475	0.9431
ACT2	0.9435	0.9462	0.9447
Overall Average	**0.9385**	**0.9407**	**0.9393**

The Precision and Recall scores further support this observation. ACL-ARC achieves an average Precision of 0.9331 and Recall of 0.9284, while SciCite obtains 0.9390 and 0.9475, respectively. ACT2 achieves the highest overall performance, with an average Precision of 0.9435, Recall of 0.9462, and F1 of 0.9447. The consistently high Precision and Recall values suggest that the generated paraphrases exhibit strong semantic correspondence with the original citation contexts from both directions: the generated texts contain substantial semantic information aligned with the originals, while the original semantic content is also well represented in the generated texts.

Importantly, the BERTScore values are high but do not approach perfect similarity. This indicates that the generated samples preserve a substantial amount of the original semantic information without being exact textual replicas. Such a property is desirable for contrastive learning, where positive samples should remain semantically aligned with the corresponding anchor while providing sufficiently different linguistic realizations. The generated paraphrases can therefore introduce additional semantic views of the same citation context and enrich the diversity of positive training instances.

Overall, the BERTScore results demonstrate that the proposed LLM augmentation strategy achieves high semantic fidelity across different citation-intent datasets. With an overall average F1 of 0.9393, the generated samples preserve substantial semantic information from their original citation contexts while maintaining sufficient linguistic variation. This balance supports their use as additional positive instances for strengthening the supervised and contrastive learning signals in TempCL-Cite.

5. Conclusions

In this work, we proposed TempCL-Cite, a temporal-aware contrastive learning framework for citation intent classification. The framework targets two limitations of existing methods: temporal information in citation relationships is rarely exploited, and high-quality annotated data for constructing informative contrastive pairs are limited. TempCL-Cite addresses these issues by jointly incorporating publication time gaps and LLM-generated intent-preserving positive examples into a unified learning framework.

TempCL-Cite uses the publication time gap Δt in two ways. It is explicitly encoded and fused with the textual representation to provide temporal information for intent prediction,

and it is also used to guide negative-sample construction by requiring different intent labels and different temporal gaps. In parallel, a prompt-based LLM augmentation strategy generates semantically consistent paraphrases of citation contexts while preserving their original citation intents. These paraphrases serve as additional positive examples during contrastive training. The LLM is used only in the offline augmentation stage and is not required during inference.

Experiments on ACL-ARC, SciCite, and ACT2 show that TempCL-Cite obtains the highest mean Accuracy and Macro-F1 among the compared methods. The largest numerical improvement appears on ACL-ARC, where TempCL-Cite reaches 82.01% Accuracy and 77.12% Macro-F1. We do not attribute the variation in improvement across datasets to dataset size alone, since class definitions, class imbalance, and data quality may also affect performance. The reported results are averaged over three independent runs; because formal significance testing was not performed, small numerical differences should be interpreted with care.

The ablation results indicate that both temporal modeling and contrastive learning contribute to performance in our experiments, and using them together gives the best mean scores. In the negative-sampling comparison, the temporal-gap constraint outperforms both label-only selection and matched-number random sampling on all three datasets, indicating that the improvement is not merely caused by using fewer negative samples. Comparisons with same-label positives and dropout views further show that LLM-generated positives achieve the highest mean scores in our setting. These comparisons support the practical value of the proposed augmentation, although they do not isolate every possible effect of additional training text.

We also evaluated the semantic similarity of the LLM-generated positives using BERTScore. The generated examples obtain an average BERTScore F1 of 0.9393 across the three datasets, indicating strong semantic correspondence with the original citation contexts while retaining linguistic variation. BERTScore measures semantic similarity rather than rhetorical function, and therefore does not by itself certify that every paraphrase preserves its citation intent. A human audit or an independent held-out classifier would provide more direct evidence and remains desirable for future work.

Overall, the results support temporal-aware contrastive learning with LLM-based augmentation as a practical approach for citation intent classification, particularly when LLM inference at deployment time is undesirable. Temporal information can serve both as an auxiliary feature and as a criterion for constructing contrastive pairs, while offline LLM augmentation provides additional semantic views without adding inference cost.

Several directions remain for future research. More fine-grained temporal representations, such as continuous publication dates, temporal decay functions, or learned temporal embeddings, could better capture the dynamics of scholarly communication; diagnostic controls such as shuffled time gaps and temporal-feature-only classifiers would help clarify the independent contribution of temporal information. The augmentation strategy could also be extended to generate multiple diverse positive views, together with automatic quality control and human or model-based verification of intent preservation. Finally, future work could examine the transferability of temporal-aware citation representations across disciplines, citation taxonomies, and heterogeneous scholarly databases, and provide more detailed class-level, error, and statistical analyses.

Author Contributions: Conceptualization, Y.S.; methodology, Y.S.; software, Y.S.; validation, X.L.; formal analysis, X.L.; investigation, Y.S. and L.W.; resources, X.L.; data curation, Y.S.; writing—original draft preparation, Y.S. and L.W.; writing—review and editing, Z.Z.; visualization, L.W.; supervision, Z.Z.; project administration, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data supporting the findings of this study are openly available on Zenodo at https://doi.org/10.5281/zenodo.23238289.

Acknowledgments: During the preparation of this manuscript/study, the authors used ChatGPT (GPT-6, OpenAI) to improve English grammar, refine sentence structure, and enhance the clarity and readability of the manuscript. In addition, the Qwen3-series model Qwen-Plus (version: qwen-plus-2025-07-14, provided by Alibaba Cloud Model Studio) was used in the proposed methodology to generate intent-preserving paraphrases of citation contexts as positive examples for contrastive learning. The authors have reviewed and edited the AI-assisted writing output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zhang, Y.; Wang, Y.; Sheng, Q.Z.; Yao, L.; Chen, H.; Wang, K.; Mahmood, A.; Zhang, W.E.; Zaib, M.; Sagar, S.; et al. Deep learning meets bibliometrics: A survey of citation function classification. *J. Inf.* **2025**, *19*, 101608. [CrossRef]
2. Sharma, K.; Uddin, Z. Beyond the h-index: Evaluating continuous research impact using the Kz index. *Scientometrics* **2026**, *131*, 3193–3217. [CrossRef]
3. Small, H. Characterizing highly cited method and non-method papers using citation contexts: The role of uncertainty. *J. Inf.* **2018**, *12*, 461–480. [CrossRef]
4. Jurgens, D.; Kumar, S.; Hoover, R.; McFarland, D.; Jurafsky, D. Measuring the evolution of a scientific field through citation frames. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 391–406. [CrossRef]
5. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543.
6. Conneau, A.; Kiela, D.; Schwenk, H.; Barrault, L.; Bordes, A. Supervised learning of universal sentence representations from natural language inference data. *arXiv* **2017**, arXiv:1705.02364.
7. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 3615–3620.
8. Paolini, L.; Vahdati, S.; Di Iorio, A.; Wardenga, R.; Heibi, I.; Peroni, S. CiteFusion: An ensemble framework for citation intent classification harnessing dual-model binary couples and SHAP analyses. *Scientometrics* **2025**, *130*, 5911–5981. [CrossRef]
9. Berrebbi, D.; Huynh, N.; Balalau, O. Graphcite: Citation intent classification in scientific publications via graph embeddings. In Proceedings of the Companion Proceedings of the Web Conference, Lyon, France, 25–29 April 2022; pp. 779–783.
10. Fogelson, A.; Trišović, A.; Thompson, N. LLMs in Citation Intent Classification: Progress, Precision, and Reproducibility Challenges. In Proceedings of the the 3rd ACM Conference on Reproducibility and Replicability, Vancouver, BC, Canada, 29–31 July 2025; pp. 250–253.
11. Karaca, K.S.; Eravcı, B. A Large-Scale Dataset and Citation Intent Classification in Turkish with LLMs. In *Proceedings of the 2025 10th International Conference on Computer Science and Engineering (UBMK)*; IEEE: Piscataway, NJ, USA, 2025; pp. 509–514.
12. Koloveas, P.; Chatzopoulos, S.; Vergoulis, T.; Tryfonopoulos, C. Can llms predict citation intent? an experimental analysis of in-context learning and fine-tuning on open llms. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries*; Springer: Berlin/Heidelberg, Germany, 2025; pp. 207–224.
13. Li, T.; Wang, J.; Zhang, Y.; Li, S.; Chen, L. Adapting Pretrained Language Models for Citation Classification via Self-Supervised Contrastive Learning. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, Toronto, ON, Canada, 3–7 August 2025; pp. 1541–1552.
14. Teufel, S.; Siddharthan, A.; Tidhar, D. Automatic classification of citation function. In Proceedings of the the 2006 Conference on Empirical Methods in Natural Language Processing, Sydney, Australia, 22–23 July 2006; pp. 103–110.
15. Abu-Jbara, A.; Ezra, J.; Radev, D. Purpose and polarity of citation: Towards nlp-based bibliometrics. In Proceedings of the he 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human language Technologies, Atlanta, GA, USA, 9–14 June 2013; pp. 596–606.
16. Hassan, S.U.; Imran, M.; Iqbal, S.; Aljohani, N.R.; Nawaz, R. Deep context of citations using machine-learning models in scholarly full-text articles. *Scientometrics* **2018**, *117*, 1645–1662. [CrossRef]

17. Cohan, A.; Ammar, W.; Van Zuylen, M.; Cady, F. Structural scaffolds for citation intent classification in scientific publications. *arXiv* **2019**, arXiv:1904.01608.
18. Yousif, A.; Niu, Z.; Chambua, J.; Khan, Z.Y. Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification. *Neurocomputing* **2019**, *335*, 195–205. [CrossRef]
19. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedingsof the the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
20. Roman, M.; Shahid, A.; Khan, S.; Koubaa, A.; Yu, L. Citation intent classification using word embedding. *IEEE Access* **2021**, *9*, 9982–9995. [CrossRef]
21. Maheshwari, H.; Singh, B.; Varma, V. SciBERT sentence representation for citation context classification. In Proceedings of the the Second Workshop on Scholarly Document Processing, Online, 10 June 2021; pp. 130–133.
22. Tsai, H.J.; Yen, A.Z.; Huang, H.H.; Chen, H.H. Citation intent classification and its supporting evidence extraction for citation graph construction. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, UK, 21–25 October 2023; pp. 2472–2481.
23. Du, X.; Ahrabian, K.; Ananthan, A.B.S.; Myloth, R.D.; Pujara, J. Graph-based structure aware citation intent classification. In Proceedings of the AAAI 2023 Workshop, Washington, DC, USA, 7–14 February 2023.
24. Qi, R.; Wei, J.; Shao, Z.; Li, Z.; Chen, H.; Sun, Y.; Li, S. Multi-task learning model for citation intent classification in scientific publications. *Scientometrics* **2023**, *128*, 6335–6355. [CrossRef]
25. Shui, Z.; Karypis, P.; Karls, D.S.; Wen, M.; Manchanda, S.; Tadmor, E.B.; Karypis, G. Fine-tuning language models on multiple datasets for citation intention classification. *arXiv* **2024**, arXiv:2410.13332.
26. Zhang, Y.; Wang, Y.; Sheng, Q.Z.; Mahmood, A.; Emma Zhang, W.; Zhao, R. TDM-CFC: Towards document-level multi-label citation function classification. In *Proceedings of the International Conference on Web Information Systems Engineering*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 363–376.
27. Kunnath, S.N.; Pride, D.; Knoth, P. Prompting strategies for citation classification. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, Birmingham, UK, 21–25 October 2023; pp. 1127–1137.
28. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. Qwen3 technical report. *arXiv* **2025**, arXiv:2505.09388.
29. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673. [CrossRef]
30. Kunnath, S.N.; Stauber, V.; Wu, R.; Pride, D.; Botev, V.; Knoth, P. ACT2: A multi-disciplinary semi-structured dataset for importance and purpose classification of citations. In Proceedings of the the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 3398–3406.
31. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. *ICLR* **2022**, *1*, 3.
32. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5753–5763.
33. Bouthillier, X.; Delaunay, P.; Bronzi, M.; Trofimov, A.; Nichyporuk, B.; Szeto, J.; Mohammadi Sepahvand, N.; Raff, E.; Madan, K.; Voleti, V.; et al. Accounting for variance in machine learning benchmarks. *Proc. Mach. Learn. Syst.* **2021**, *3*, 747–769.
34. Wang, S. Automated fault detection and diagnosis of ahus via tabular-based methods using operational data from a large office building. *J. Comput. Civ. Eng.* **2026**, *40*, 04026018. [CrossRef]
35. Wang, S. Fault-class coverage–aligned combined training for AFDD of AHUs across multiple buildings. *Sci. Rep.* **2025**, *15*, 41192. [CrossRef]
36. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. Bertscore: Evaluating text generation with bert. *arXiv* **2019**, arXiv:1904.09675.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.