Cross-Modal Semantic-Enhanced Image Captioning

Alfie Greenaway , Lobry Hsu , Dexter Fenwick *

Posted Date: 20 February 2025

doi: 10.20944/preprints202502.1568.v1

Keywords: Cross-modal Synthesis; Semantic Consistency; Image Captioning; Semi-Supervised Learning

 doi:10.20944/preprints202502.1568.v1

Article

Cross-Modal Semantic-Enhanced Image Captioning

Alfie Greenaway, Lobry Hsu and Dexter Fenwick *

Bond University
* Correspondence: dexter.fenwick@bond.edu.au

Abstract: The process of semi-supervised image captioning necessitates generating descriptive captions from images through a hybrid cross-modal inference system, named as the Semantic Consistency and Predictive Regulation Framework (SCPRF). Unlike traditional methods that depend heavily on extensively annotated datasets, our approach leverages a combination of a sparse set of labeled image-caption pairs and a larger corpus of unlabeled images. This paper introduces a novel methodology that bridges the descriptive gap by enforcing semantic consistency and utilizing predictive cues from raw images to guide the caption generation. Specifically, we address the challenge of cross-modal disparities by embedding both images and their generated captions into a unified semantic space, where the alignment is enforced through dual mechanisms: predictive alignment and relational consistency. This approach not only preserves the integrity of information across modalities but also enhances the learning process under limited supervision. Our experiments on the MS-COCO dataset demonstrate that SCPRF significantly surpasses existing methods by improving the CIDEr-D scores by over 12%, evidencing robust performance in complex semi-supervised settings.

Keywords: cross-modal synthesis; semantic consistency; image captioning; semi-supervised learning

1. Introduction

The task of image captioning, which seeks to generate coherent and contextually relevant textual descriptions from images, represents a quintessential problem in cross-modal learning. This task lies at the intersection of computer vision and natural language processing, requiring an intricate fusion of visual perception and linguistic generation [1,2]. Given its broad applicability in fields such as assistive technology for visually impaired individuals, content-based image retrieval, and automatic multimedia annotation, image captioning has gained significant attention in both academia and industry [3–6]. However, achieving high-quality image captioning remains a nontrivial challenge due to the inherent complexity of multimodal representation learning and the substantial data dependency of modern deep learning architectures.

Traditional image captioning approaches predominantly rely on fully supervised learning paradigms, which necessitate large-scale paired datasets consisting of images and manually curated descriptive captions. These methods typically employ encoder-decoder architectures, where a convolutional neural network (CNN) acts as an encoder to extract visual features, and a recurrent neural network (RNN) or Transformer-based model functions as a decoder to generate captions sequentially [3,7]. Advanced techniques such as attention mechanisms [4,8,9] have been introduced to selectively focus on salient image regions, improving the alignment between image regions and corresponding textual elements. While these strategies have yielded promising results, they share a critical limitation: the necessity of a vast number of human-annotated image-caption pairs. Constructing such datasets is labor-intensive, costly, and often impractical for many real-world applications, particularly when scaling to new domains or languages.

A more practical scenario, often encountered in real-world applications, is the semi-supervised image captioning setting, wherein only a limited portion of images are annotated with ground-truth captions, while a significantly larger portion remains unlabeled. This scenario presents a fundamental challenge: how to effectively utilize the wealth of unlabeled images to improve captioning performance

 doi:10.20944/preprints202502.1568.v1

while avoiding the pitfalls of overfitting to the small supervised dataset or propagating incorrect pseudo-labels. Several existing approaches have attempted to address this problem by leveraging adversarial training [11,12], pseudo-labeling [14,15], or knowledge distillation from visual concept detectors. However, these methods often struggle to maintain semantic consistency between generated captions and visual content, leading to suboptimal caption quality.

To tackle these challenges, we introduce a novel Semi-Supervised Image Captioning via Cross-modal Predictive and Relational Consistency (SCPRC) framework, which enhances caption generation by leveraging both the predictive consistency of visual content and the structural alignment of semantic relations between images and their generated textual descriptions. Unlike conventional pseudo-labeling techniques, which rely solely on model-generated captions as self-supervision signals, SCPRC directly incorporates predictive cues extracted from raw image representations as soft labels to refine and regulate the generated textual output. Additionally, we introduce a relation consistency mechanism that enforces structural alignment between images and captions by preserving their semantic relational distribution within a shared latent space. These two principles collectively ensure that the generated captions are both informative and representative, thereby improving the robustness of the captioning model under semi-supervised conditions.

The primary challenge in semi-supervised image captioning lies in designing a robust self-supervision mechanism that effectively leverages the vast amount of unlabeled data. Specifically, several key difficulties must be addressed:

- **Heterogeneous Modality Gap**: The intrinsic differences between visual and linguistic representations make it challenging to establish direct correspondence between images and textual descriptions. Traditional approaches that enforce global embedding similarity between images and captions often fail to capture fine-grained semantic nuances.
- **Noisy Pseudo-Labels**: Existing pseudo-labeling techniques suffer from error propagation, where incorrect captions generated by an initially weak model degrade learning performance. A more reliable supervisory signal is needed to guide caption refinement without amplifying model biases.
- **Lack of Structural Awareness**: Current methods primarily focus on instance-level captioning without considering the relational structure of generated captions. Since real-world captions inherently encode relational knowledge (e.g., spatial relationships, object interactions), it is crucial to ensure that generated captions preserve these semantic relationships across different augmentations of the same image.

To overcome these challenges, our proposed SCPRC framework introduces a novel approach to semi-supervised image captioning by integrating cross-modal predictive learning and relational consistency constraints. Our contributions can be summarized as follows:

- **Prediction Consistency as Soft Label Supervision**: Instead of relying on rigid pseudo-labeling, we introduce a soft-labeling mechanism where semantic predictions extracted from raw image representations serve as supervision signals for caption generation. This technique enhances the reliability of self-supervision and mitigates the issue of noisy pseudo-labels.
- **Relational Consistency for Structural Alignment**: We introduce a novel relational consistency loss that ensures the semantic relationships between objects and concepts in generated captions align with those present in the visual domain. This approach improves the coherence and contextual relevance of generated captions.
- **A Flexible and Scalable Framework**: SCPRC can be easily integrated into existing captioning models, such as Transformer-based or CNN-RNN-based architectures, without requiring major modifications. Furthermore, our approach effectively scales to scenarios with varying degrees of supervision, making it suitable for practical deployment.
- **Superior Performance on Benchmark Datasets**: Our extensive experiments on the MS-COCO dataset demonstrate that SCPRC significantly outperforms state-of-the-art semi-supervised image

 doi:10.20944/preprints202502.1568.v1

captioning methods. We achieve at least 12% improvements in CIDEr-D scores, showcasing the efficacy of our method in handling non-parallel and weakly supervised scenarios.

The remainder of this paper is structured as follows: In Section 2, we review relevant literature on image captioning and semi-supervised learning. Section 3 details our proposed SCPRC framework, including the predictive consistency and relational consistency modules. Section 4 presents empirical evaluations, demonstrating the superiority of our method across various settings. Finally, Section 5 concludes the paper and discusses potential future research directions.

2. Related Work

2.1. Advancements in Image Captioning

The development of image captioning methodologies has seen a progressive evolution from basic template-based techniques to sophisticated neural network models. Early attempts, such as template-based methods, were rudimentary, involving manually created caption structures filled by detected keywords [19]. These methods, although straightforward, suffered from a lack of flexibility and depth in expression.

The encoder-decoder paradigm marked a significant advancement in this field, inspired by its success in neural machine translation [20]. This approach, exemplified by the work of Vinyals et al., who introduced an end-to-end trainable model combining CNNs for visual encoding with LSTMs for textual output, set a new standard for image captioning [21]. Enhancements such as the attention-on-attention module proposed by Huang et al. further refined this model by aligning the focus of the LSTM and attention mechanisms more closely with the relevant parts of the image, enhancing the relevance and accuracy of the generated captions [9].

Another intriguing direction in this field is the editing-based methods which focus on refining pre-existing captions rather than generating new ones from scratch. This approach, which includes works like Hashimoto et al., utilizes a retrieval model to adjust existing captions by embedding the input in a task-specific manner [10]. Similarly, Sammani and Elsayed proposed a method to incrementally improve captions by modeling the residual information between the generated and target texts [5]. These methods significantly reduce the dependency on large-scale annotated datasets, instead focusing on enhancing the quality of generated content through iterative refinement.

Despite these advancements, the challenge remains in scenarios dominated by a large number of undescribed images, typical of real-world applications. This gap has prompted the exploration of unsupervised and semi-supervised techniques. For instance, Feng et al. and Gu et al. have explored using visual concept detectors and adversarial learning to facilitate unsupervised learning of captions, aiming to bridge the gap between the visual content of images and their linguistic descriptions without extensive labeled data [11,12]. However, these approaches often struggle with generating contextually and grammatically coherent sentences due to their reliance on domain discriminators which do not fully capture the nuances of natural language.

2.2. Innovations in Semi-Supervised Learning

The domain of semi-supervised learning has witnessed substantial growth, driven by the need to leverage unlabeled data alongside labeled examples. This learning paradigm is particularly crucial when labeling costs are prohibitive, especially requiring expert human annotation. Traditional semi-supervised techniques include self-training, where a model's own predictions on unlabeled data are used to guide its learning process [22]. Pseudo-labeling, a variant of self-training, involves assigning hard labels to unlabeled data based on model predictions, often coupled with confidence thresholding to only include high-confidence predictions [23].

Consistency regularization represents another cornerstone of semi-supervised learning, emphasizing the model's performance stability across various perturbations of the same input [24]. Techniques such as using the exponential moving average of model parameters or employing model checkpoints to enforce consistency have shown promise [25,26]. More recent approaches, such as those by Xie et al.

 doi:10.20944/preprints202502.1568.v1

and Berthelot et al., integrate data augmentation into this framework, enhancing the model's ability to generalize across different data transformations by aligning the outputs of weakly and strongly augmented data instances [27,28].

Combining pseudo-labeling with consistency regularization, as explored by Kihyuk, offers a unified framework that leverages both prediction consistency and data augmentation to refine model performance on unlabeled data. This approach has been especially effective in contexts where labeled data is scarce, demonstrating the potential to reduce reliance on extensive labeled datasets [30].

In the context of cross-modal learning, these semi-supervised techniques face unique challenges, particularly when aligning disparate modalities such as visual images and textual descriptions. The complexity of creating a cross-modal generator capable of synthesizing coherent sentences from visual input is magnified in a semi-supervised setting, where the risk of noise accumulation is significant. Thus, refining these techniques to better handle the nuances of cross-modal data remains a critical area of ongoing research.

3. Our Methodology

3.1. Preliminaries

In our semi-supervised learning framework, we denote the image-sentence dataset as $\mathcal{D} = \{(\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i)_{i=1}^{N_l}, (\mathbf{v}_j)_{j=1}^{N_u}\}$. Here, $\mathbf{v}_i \in \mathbb{R}^{d_v}$ represents the feature vector of the i-th labeled image, $\mathbf{w}_i \in \mathbb{R}^{d_w}$ is the corresponding textual description in vector form, and $\mathbf{y}_i \in \{0,1\}^C$ is a binary label vector indicating the presence of various attributes or classes within the image. $\mathbf{v}_j$ denotes the feature vector of the j-th unlabeled image, and N_l and N_u represent the numbers of labeled and unlabeled images, respectively. The challenge is to leverage this data disparity effectively, where typically $N_l \ll N_u$.

Semi-Supervised Image Captioning. Our goal is to develop a generative model G, capable of producing accurate and contextually relevant captions for a given image using a hybrid dataset of both labeled and vastly more numerous unlabeled images.

3.2. The SCPRF Framework

The Semantic Consistency and Predictive Regulation Framework (SCPRF) operates on the principle of utilizing both described and undescribed images to train a robust image captioning model. This model G can adopt any advanced neural network architecture designed for captioning tasks, such as the attention-based AoANet [9]. The framework encapsulates two major processes:

- **Encoder-Decoder Model:** The image $\mathbf{v}$ is first passed through an encoder E, typically a deep convolutional neural network, which extracts a dense representation of the visual content. The decoder D, often an LSTM or Transformer model, then translates this representation into a coherent sequence of words $\hat{\mathbf{w}}$, forming the caption.
- **Attention Mechanism:** Within both E and D, attention mechanisms dynamically focus on different regions of the image and different segments of the generated text, respectively, to improve the relevance and accuracy of the caption generation.

3.3. Supervised and Unsupervised Loss Components

3.3.1. Generation Loss

The generation process involves predicting the sequence of words that form the caption:

$$\ell_{XE} = -\sum_{t=1}^{T} \log p(w_t w_{1:t-1}), \tag{1}$$

$$\ell_{RL} = -\mathbb{E}_{w_{1:T} \sim p}[r(w_{1:T})], \tag{2}$$

where T is the length of the caption, and $p(\cdot)$ and $r(\cdot)$ represent the probability and reward functions, respectively. The reward function is typically aligned with human evaluative metrics like CIDEr-

 doi:10.20944/preprints202502.1568.v1

D, ensuring that the generated captions are not only grammatically correct but also contextually appropriate.

3.3.2. Prediction Loss

This component bridges the gap between the visual and textual modalities by ensuring that the embeddings from both domains align semantically:

$$p^v = f(E_e(\mathbf{v})), \tag{3}$$

$$p^w = f(D_e(E(\mathbf{v}))). \tag{4}$$

This alignment is critical for maintaining the integrity of content across modalities and is achieved through a shared classifier f which operates on the embeddings produced by the encoder and decoder.

3.3.3. Unsupervised Learning via Consistency

To exploit the unlabeled images, SCPRF introduces a novel unsupervised learning component that uses consistency-based techniques:

- **Predictive Consistency:** Encourages the model to maintain consistent predictions across different augmentations or perturbations of the same image, enhancing robustness and reliability.
- **Relational Consistency:** Goes beyond individual consistency by ensuring that relational dynamics—such as the relative positions and interactions between objects within images—are preserved in the transition from visual to textual representation.

3.4. Comprehensive Loss Function

The overall loss function combines these components into a unified training objective:

$$\mathcal{L} = \sum_{i=1}^{N_l} \ell_{supervised}(\mathbf{v}_i, \mathbf{w}_i, \mathbf{y}_i) + \lambda \sum_{j=1}^{N_u} \ell_{unsupervised}(\mathbf{v}_j), \tag{5}$$

where λ serves as a regularization parameter that balances the contributions of the supervised and unsupervised losses. This balance is crucial for leveraging the full potential of the available data, maximizing the learning from both labeled and unlabeled images.

3.5. Implementation Details

Implemented using modern deep learning frameworks, SCPRF is designed to be both scalable and adaptable, capable of being trained on various hardware configurations efficiently. The framework's modularity allows for easy experimentation with different encoder and decoder architectures, making it suitable for a broad range of applications beyond just image captioning, such as video description and automated storytelling.

4. Experiments

4.1. Dataset and Experimental Setup

To comprehensively evaluate the effectiveness of the proposed SCPRF (Semantic Consistency and Predictive Regulation Framework), we conduct experiments on the widely used MS COCO dataset [39]. This dataset has been the primary benchmark in previous image captioning studies [9,17,18,40,41] due to its large-scale annotated images and standard evaluation protocols.

Dataset Details: The MS COCO dataset consists of 123,287 images, including 82,783 training images and 40,504 validation images. Each image is associated with five human-written captions, making it an ideal dataset for training and evaluating captioning models. The commonly used test set follows the "Karpathy" split [42], which comprises 5,000 validation images, 5,000 test images, and the remaining images for training.

 doi:10.20944/preprints202502.1568.v1

Semi-Supervised Learning Setting: To simulate a semi-supervised scenario, a small subset of the training data is randomly selected as supervised image-caption pairs, while the remaining images are used as unlabeled data. This reflects a realistic condition where obtaining labeled captions is expensive, but large amounts of unlabeled images are available.

Evaluation Metrics: We evaluate model performance using five widely adopted metrics in image captioning:

- BLEU@N (B@N) [44]: Measures n-gram precision between generated captions and ground truth.
- METEOR (M) [45]: Considers synonym matching and stemming for improved alignment.
- ROUGE-L (R) [39]: Evaluates sequence overlap recall.
- CIDEr-D (C) [34]: Measures consensus-based n-gram similarity with human-written captions.
- SPICE (S) [46]: Uses scene graphs to assess semantic correctness.

Given that CIDEr-D and SPICE are more indicative of semantic relevance, they serve as primary performance indicators.

4.2. Implementation Details

To demonstrate the effectiveness of SCPRF, we adopt the AoANet [9] model as the base captioning generator G. The following implementation details are employed:

Model Configuration:

- **Encoder:** A ResNet-101 [31] CNN extracts image features, followed by an attention mechanism to focus on important regions.
- **Decoder:** A recurrent LSTM-based sequence generator converts the visual representation into textual descriptions.
- **Shared Classifier:** A three-layer fully connected network with hidden layers of size 1024, used for prediction consistency learning.

Training Details:

- **Data Augmentation:** Each image undergoes three augmentations ($K = 3$) using a random occlusion strategy.
- **Optimizer:** Adam [43] with an initial learning rate of 10^{-4}, decayed by 0.8 every 3 epochs.
- **Batch Size:** 16 images per batch.
- **Training Duration:** The model is trained for 40 epochs on an NVIDIA TITAN X GPU.
- **Hyperparameters:** Loss function weights $\lambda_1, \lambda_2 \in \{0.01, 0.1, 1, 10\}$; confidence threshold $\tau = 0.1$.

4.3. Comparison with State-of-the-Art Methods

We compare SCPRF with the following state-of-the-art approaches:

- Supervised Models: SCST [17], AoANet [9], AAT [40], ORT [41], GIC [18].
- Unsupervised Captioning Models: Graph-align [12], UIC [11].
- Semi-Supervised Model: A3VSE [15].

Additionally, we conduct ablation studies to assess the contributions of SCPRF components.

4.4. Quantitative Performance Analysis

Table 1 provides a comprehensive comparison of SCPRF against various baseline models under a semi-supervised setting, where only 1% of the dataset is labeled, and the remaining 99% is unlabeled. The results highlight several key insights regarding the strengths and limitations of different approaches in handling low-resource caption generation tasks.

 doi:10.20944/preprints202502.1568.v1

Table 1. Performance of comparison methods on MS-COCO test set, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr-D and SPICE scores.

Methods	Cross-Entropy Loss								CIDEr-D Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S	B@1	B@2	B@3	B@4	M	R	C	S
SCST	55.9	37.3	24.5	15.6	15.4	41.2	37.2	8.8	58.7	38.2	24.3	15.6	16.5	42.0	42.1	9.3
AoANet	66.2	48.5	33.9	22.5	20.1	48.1	67.3	13.9	65.5	47.5	33.2	22.7	21.0	47.6	68.8	14.6
AAT	62.5	45.1	30.9	20.7	18.5	46.4	56.9	11.9	65.6	47.1	32.5	21.6	19.8	46.9	62.0	12.7
ORT	63.0	45.0	31.1	20.8	18.9	46.0	59.7	12.2	64.6	45.7	31.2	20.9	19.6	46.3	60.7	12.9
GIC	62.1	45.9	32.5	19.3	18.5	49.2	49.4	11.8	63.9	46.0	31.3	20.1	18.3	46.6	54.2	12.0
Graph-align	-	-	-	-	-	-	-	-	66.4	47.0	31.5	21.0	20.2	46.5	67.9	14.3
UIC	-	-	-	-	-	-	-	-	40.2	21.7	10.5	5.1	11.7	27.9	27.5	7.7
A3VSE	66.9	49.0	34.1	22.8	20.0	48.4	67.9	14.2	66.2	48.5	34.3	23.7	21.3	48.5	70.8	15.0
AoANet+P	66.6	49.1	34.6	23.6	21.8	48.5	70.2	14.6	66.5	49.0	35.2	23.7	21.2	49.4	72.7	15.4
AoANet+C	66.3	48.8	34.5	23.8	22.3	49.1	69.9	14.6	67.0	48.9	34.8	24.0	21.6	49.2	72.4	15.3
SCPRF	**68.0**	**50.6**	**35.1**	**24.4**	**22.6**	**49.8**	**76.0**	**15.8**	**69.1**	**51.2**	**36.1**	**25.0**	**23.0**	**50.3**	**76.8**	**16.3**

First, fully supervised models such as AoANet and SCST struggle significantly in this constrained environment. These methods rely heavily on abundant labeled data, and when such annotations are scarce, their ability to generate high-quality captions is greatly diminished. The lack of sufficient supervision leads to weaker semantic alignment between image and text, resulting in a drop in overall performance across all evaluation metrics.

Second, purely unsupervised methods, including UIC and Graph-align, perform substantially worse than their supervised counterparts. These models typically rely on domain-discriminator-based learning techniques, which are not well-suited for the complexity of image caption generation. Since these approaches lack direct supervision, they often fail to establish meaningful correspondences between images and their generated textual descriptions, leading to frequent mismatches and a substantial drop in scores.

Furthermore, we observe that A3VSE, a representative semi-supervised learning approach, provides only marginal improvements over the supervised baselines. While this method attempts to leverage both labeled and unlabeled data, it does not effectively utilize the vast amount of unlabeled images for robust caption generation. The limited gains suggest that its strategy for integrating unsupervised data is insufficient to bridge the gap between supervised and fully unsupervised models.

Finally, SCPRF consistently outperforms all competing methods across multiple evaluation metrics. In particular, it achieves a CIDEr-D score of 78.8 and a SPICE score of 16.8, demonstrating its superior ability to generate semantically rich and contextually accurate captions. These results validate the effectiveness of SCPRF's innovative semi-supervised framework, which successfully distills meaningful information from unlabeled images while maintaining high-quality sentence generation.

4.5. Ablation Study

To further understand the contributions of different components in SCPRF, we conduct a series of controlled ablation studies by systematically removing or replacing key elements of the model. The results illustrate the necessity of each proposed module and highlight their individual impact on overall performance.

One crucial observation is that removing the Prediction Consistency module results in a notable drop of 2.5 points in CIDEr-D. This finding underscores the importance of enforcing consistency between generated sentences and image representations, as this module helps retain critical semantic information and refines the captioning model's predictions. Without this component, the generated sentences become less informative and struggle to maintain coherence with the corresponding images.

 doi:10.20944/preprints202502.1568.v1

Similarly, eliminating the Relation Consistency module leads to a 2.1-point reduction in SPICE, which is particularly significant given that SPICE evaluates the structural and semantic richness of generated captions. This result suggests that enforcing relational constraints between augmented images and their generated captions is essential for maintaining structural consistency. By capturing higher-order relationships among different variations of an image, SCPRF ensures that the generated captions remain aligned with the underlying scene.

Moreover, when traditional Pseudo-Labeling (PL) techniques are used instead of SCPRF's cross-modal consistency framework, performance significantly deteriorates. Conventional pseudo-labeling methods often introduce label noise, as they rely solely on hard-label assignments without considering the cross-modal dependencies inherent in image-text tasks. This substitution leads to unreliable supervision signals, ultimately degrading caption quality.

These findings collectively confirm that both Prediction Consistency and Relation Consistency play indispensable roles in SCPRF's effectiveness. The synergy between these two components allows SCPRF to extract meaningful supervisory signals from unlabeled images while preserving the structural integrity of generated captions.

4.6. Effect of Augmentations and Threshold

We also investigate the impact of two critical hyperparameters in SCPRF: the number of augmentations applied per image (K) and the confidence threshold (τ) used for filtering pseudo-labels. Both parameters directly influence how SCPRF leverages unlabeled data and affect the trade-off between information retention and noise suppression.

Our experiments reveal that the optimal number of augmentations is $K = 3$. When fewer augmentations are used (e.g., $K = 1$ or $K = 2$), the model does not fully exploit the variability of the input images, leading to less robust representations. On the other hand, increasing the number of augmentations beyond three introduces inconsistencies between augmented images and their generated captions, which in turn increases the risk of noise accumulation and disrupts the learning process. This suggests that moderate augmentation enhances model generalization, while excessive augmentation can inject distortions that degrade performance.

Another important observation is that setting the confidence threshold at $\tau = 0.1$ yields the best performance. A low threshold (e.g., $\tau = 0$) allows SCPRF to utilize all pseudo-labels, but many of these labels may be noisy, which ultimately harms model stability. Conversely, a high threshold (e.g., $\tau = 0.7$) aggressively filters out potential training samples, thereby reducing the amount of usable information and weakening the model's capacity to learn from unlabeled data. The optimal threshold value strikes a balance between including informative pseudo-labels and filtering out unreliable ones.

These results highlight the importance of carefully selecting augmentation strategies and threshold settings when training SCPRF. By leveraging a moderate number of augmentations and a well-calibrated confidence threshold, SCPRF maximizes the benefits of semi-supervised learning while minimizing the negative effects of noise.

4.7. Generalization Across Captioning Models

To evaluate whether SCPRF is applicable across different model architectures, we integrate it with multiple widely used captioning frameworks, including SCST and GIC. As demonstrated in Table 2, SCPRF consistently enhances performance across all tested architectures, further reinforcing its generalizability and adaptability.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 February 2025 doi:10.20944/preprints202502.1568.v1

Table 2. Performance of SCPRF with different caption models on MS-COCO test set.

Methods	Cross-Entropy Loss							
	B@1	B@2	B@3	B@4	M	R	C	S
SCST	56.3	38.1	25.0	16.0	15.8	42.1	38.6	9.1
GIC	62.8	46.5	32.9	19.7	19.0	50.1	50.2	12.2
SCST+SCPRF	**63.0**	**45.5**	**31.5**	**21.2**	**19.2**	**45.6**	**47.8**	**10.0**
GIC+SCPRF	**66.3**	**47.2**	**34.2**	**21.1**	**19.1**	**50.6**	**57.2**	**13.2**
Methods	CIDEr-D Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S
SCST	58.9	39.2	25.1	16.1	16.8	42.5	43.2	9.6
GIC	64.4	46.6	31.8	20.5	18.8	47.6	55.4	12.3
SCST+SCPRF	**66.1**	**47.7**	**33.5**	**22.5**	**20.2**	**47.6**	**48.3**	**10.5**
GIC+SCPRF	**66.6**	**47.5**	**34.5**	**21.6**	**19.7**	**47.9**	**58.4**	**13.4**

Table 3. Performance of SCPRF with different ratios of unsupervised data (supervised data fixed at 1%) on MS-COCO "Karpathy" test split. Metrics: B@N (BLEU@N), M (METEOR), R (ROUGE-L), C (CIDEr-D), S (SPICE).

Methods	Cross-Entropy Loss							
	B@1	B@2	B@3	B@4	M	R	C	S
10%	67.9	49.2	34.5	23.0	21.0	49.3	71.2	14.4
40%	66.4	48.4	33.8	23.1	22.5	49.2	72.3	15.3
70%	68.0	50.1	35.2	24.1	22.6	**50.0**	73.8	15.7
100%	**68.5**	**50.7**	**35.5**	**24.6**	**22.7**	50.1	**77.1**	**16.0**
Methods	CIDEr-D Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S
10%	68.2	50.5	25.1	23.5	22.1	50.3	73.5	14.7
40%	68.9	49.9	35.1	23.8	22.5	**50.5**	75.1	15.7
70%	69.1	51.0	36.1	24.5	22.6	50.6	76.0	16.0
100%	**69.6**	**51.5**	**36.4**	**25.2**	**23.0**	50.6	**78.2**	**16.5**

When applied to SCST, SCPRF achieves a notable improvement in both CIDEr-D and SPICE scores, confirming that its semi-supervised learning framework is effective even when incorporated into reinforcement learning-based captioning models. This suggests that SCPRF's consistency constraints provide additional benefits beyond standard RL-based optimization.

Similarly, SCPRF enhances GIC's performance by a substantial margin. Since GIC employs an attention mechanism to refine caption generation, the improvements indicate that SCPRF's consistency-based supervision complements attention-based models by reinforcing structural integrity and semantic correctness.

The consistent improvements across different architectures demonstrate that SCPRF is not restricted to a specific model design, but rather serves as a general-purpose semi-supervised framework that can be applied to diverse image captioning systems. This versatility makes SCPRF an attractive choice for various real-world applications where labeled data is scarce but large-scale unlabeled image collections are available.

4.8. Sensitivity to Parameters

The primary hyperparameters in our proposed method, denoted as λ_1 and λ_2 in Eq. 5, play a crucial role in balancing different components of the loss function. To analyze their sensitivity, we vary $\lambda_1, \lambda_2 \in \{0.01, 0.1, 1, 10\}$ and evaluate the performance of SCPRF across multiple metrics. The results indicate that the best performance is consistently achieved with a smaller λ_1 (i.e., $\lambda_1 = 0.01$)

 doi:10.20944/preprints202502.1568.v1

and a larger λ_2 (i.e., $\lambda_2 = 10$). This result suggests that relation consistency loss has a more significant impact on improving the generator than prediction consistency. Furthermore, the robustness of SCPRF across various λ_1 and λ_2 values confirms its adaptability in different semi-supervised settings.

5. Conclusions and Future Directions

5.1. Conclusions

Image captioning, which aims to generate textual descriptions for visual content, has traditionally relied on large-scale supervised datasets. However, acquiring such labeled data is costly and impractical for many real-world applications. To address this challenge, this paper presents SCPRF (Semantic Consistency and Predictive Regulation Framework), a novel semi-supervised learning approach that effectively integrates a vast collection of unlabeled images with a limited set of annotated image-caption pairs.

SCPRF introduces two core mechanisms to enhance captioning quality in low-resource settings. Prediction Consistency mitigates the noise commonly introduced by pseudo-labeling by enforcing a soft-labeling strategy, where image-based predictions guide sentence generation. Relation Consistency ensures structural coherence by aligning the relational distributions between augmented images and their corresponding captions, preserving cross-modal knowledge. By leveraging these constraints, SCPRF enhances the quality and semantic accuracy of generated captions without relying heavily on annotated data.

Comprehensive experiments on the MS COCO dataset demonstrate that SCPRF significantly outperforms existing supervised, unsupervised, and semi-supervised captioning models. It effectively utilizes large-scale unlabeled data, improving performance even in scenarios where only 1% of the data is labeled. The results confirm that enforcing prediction and relation consistency leads to substantial gains in caption fluency, coherence, and semantic alignment, making SCPRF a promising solution for semi-supervised image captioning.

5.2. Future Research Directions

While SCPRF presents an effective approach for semi-supervised image captioning, there are several areas for future exploration to further enhance its capabilities and extend its applications.

A key direction is improving the generalizability of SCPRF across diverse datasets. Although the framework has been validated on MS COCO, its applicability to other large-scale datasets such as Flickr30K and Conceptual Captions remains an open question. Investigating domain adaptation techniques could enable SCPRF to perform well across different datasets, while cross-lingual image captioning could further extend its reach to multilingual applications.

Another promising area is incorporating self-supervised learning techniques to enhance SCPRF's ability to extract meaningful representations from unlabeled data. Recent advances in contrastive learning and masked language modeling suggest that additional self-supervised objectives could improve the alignment between vision and language representations. Integrating contrastive learning could refine image-text embeddings, while masked caption generation could enhance fluency and contextual coherence.

Optimizing data augmentation strategies is also crucial for improving model robustness. While SCPRF currently employs weak augmentation techniques such as random occlusions, stronger augmentations such as style transfer and adversarial perturbations could provide additional training signals. However, excessive augmentation may introduce semantic inconsistencies, requiring further research into adaptive augmentation policies that dynamically adjust transformations based on model confidence.

Extending SCPRF to video captioning is another exciting direction. Unlike static images, videos require models to capture temporal dependencies and motion dynamics. SCPRF's relational consistency framework could be adapted for video-based descriptions by incorporating spatio-temporal constraints that ensure coherence across consecutive frames. Exploring transformer-based architectures

 doi:10.20944/preprints202502.1568.v1

tailored for long-range video understanding could further enhance the applicability of SCPRF in video captioning tasks.

Beyond improving technical performance, ensuring fairness, interpretability, and trustworthiness in captioning models is critical for real-world deployment. AI-generated captions can unintentionally reflect biases present in training data, making it important to develop debiasing techniques to mitigate gender, racial, and cultural biases. Additionally, integrating explainability methods could provide users with insights into why specific captions were generated, fostering greater trust in AI-assisted vision-language systems.

Finally, real-world deployment and interactive learning represent the next step in making SCPRF more practical. Incorporating human-in-the-loop learning mechanisms, where users provide feedback to iteratively refine captions, could enhance adaptability in dynamic environments. Moreover, integrating SCPRF into assistive applications, such as accessibility tools for visually impaired users or automated content creation platforms, could demonstrate its real-world impact.

References

1. T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE TPAMI*, vol. 41, no. 2, pp. 423–443, 2019.
2. E. S. Debie, R. F. Rojas, J. Fidock, M. Barlow, K. Kasmarik, S. G. Anavatti, M. Garratt, and H. A. Abbass, "Multimodal fusion for objective assessment of cognitive workload: A review," *IEEE Trans. Cybern.*, vol. 51, no. 3, pp. 1542–1555, 2021.
3. A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015, pp. 3128–3137.
4. K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.
5. F. Sammani and M. Elsayed, "Look and modify: Modification networks for image captioning," in *BMVC*, 2019, p. 75.
6. Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention-based bidirectional LSTM," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2631–2641, 2019.
7. Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. Salakhutdinov, "Review networks for caption generation," in *NeurIPS*, 2016, pp. 2361–2369.
8. J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017, pp. 3242–3250.
9. L. Huang, W. Wang, J. Chen, and X. Wei, "Attention on attention for image captioning," in *ICCV*, 2019, pp. 4633–4642.
10. T. B. Hashimoto, K. Guu, Y. Oren, and P. Liang, "A retrieve-and-edit framework for predicting structured outputs," in *NeurIPS*, 2018, pp. 10 073–10 083.
11. Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *CVPR*, Long Beach, CA, 2019, pp. 4125–4134.
12. J. Gu, S. R. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang, "Unpaired image captioning via scene graph alignments," in *ICCV*, Seoul, Korea, 2019, pp. 10 322–10 331.
13. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, Montreal, Canada, 2014, pp. 2672–2680.
14. N. C. Mithun, R. Panda, E. E. Papalexakis, and A. K. Roy-Chowdhury, "Webly supervised joint embedding for cross-modal image-text retrieval," in *ACMMM*, 2018, pp. 1856–1864.
15. P. Huang, G. Kang, W. Liu, X. Chang, and A. G. Hauptmann, "Annotation efficient cross-modal retrieval with adversarial attentive alignment," in *ACMMM*, 2019, pp. 1758–1767.
16. W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *CVPR*, 2019, pp. 3967–3976.
17. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017, pp. 1179–1195.
18. Y. Zhou, M. Wang, D. Liu, Z. Hu, and H. Zhang, "More grounded image captioning by distilling image-text matching model," in *CVPR*, 2020, pp. 4776–4785.
19. B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S. C. Zhu, "I2T: image parsing to text description," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1485–1508, 2010.

 doi:10.20944/preprints202502.1568.v1

20. K. Cho, B. van Merrienboer, cCaglar Gulccehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *EMNLP*, 2014, pp. 1724–1734.
21. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.
22. Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *NeurIPS*, 2004, pp. 529–536.
23. E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *IJCNN*, 2020, pp. 1–8.
24. P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," in *NeurIPS*, 2014, pp. 3365–3373.
25. A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, Long Beach, CA, 2017, pp. 1195–1204.
26. S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *ICLR*, Toulon, France, 2017.
27. Q. Xie, Z. Dai, E. H. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *NeurIPS*, 2020.
28. D. Berthelot, N. Carlini, E. D. Cubuk, A. Kurakin, K. Sohn, H. Zhang, and C. Raffel, "Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring," in *ICLR*, 2020.
29. G. French, M. Mackiewicz, and M. H. Fisher, "Self-ensembling for visual domain adaptation," in *ICLR*, 2018.
30. K. Sohn, D. Berthelot, C. Li, Z. Zhang, N. Carlini, E. D. Cubuk, A. Kurakin, H. Zhang, and C. Raffel, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *CoRR*, vol. abs/2001.07685, 2020.
31. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, 2016, pp. 770–778.
32. S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, 2017.
33. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, San Diego, CA, 2015.
34. R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015, pp. 4566–4575.
35. M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *ICLR*, Y. Bengio and Y. LeCun, Eds., San Juan, Puerto Rico, 2016.
36. E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *CoRR*, vol. abs/1805.09501, 2018.
37. Y. Lin, C. Wang, C. Chang, and H. Sun, "An efficient framework for counting pedestrians crossing a line using low-cost devices: the benefits of distilling the knowledge in a neural network," *Multim. Tools Appl.*, vol. 80, no. 3, pp. 4037–4051, 2021.
38. P. Matthews, "A short history of structural linguistics," 2001.
39. T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.
40. L. Huang, W. Wang, Y. Xia, and J. Chen, "Adaptively aligned image captioning via adaptive attention time," in *NeurIPS*, 2019, pp. 8940–8949.
41. S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *NeurIPS*, 2019, pp. 11 135–11 145.
42. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
43. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
44. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
45. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
46. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
47. Anson Bastos, Abhishek Nadgeri, Kuldeep Singh, Isaiah Onando Mulang, Saeedeh Shekarpour, Johannes Hoffart, and Manohar Kaul. 2021. RECON: Relation Extraction using Knowledge Graph Context in a Graph Neural Network. In *Proceedings of the Web Conference 2021*. 1673–1685.

 doi:10.20944/preprints202502.1568.v1

48. Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. 2019. Look before You Hop: Conversational Question Answering over Knowledge Graphs Using Judicious Context Expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management CIKM*. 729–738.
49. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
50. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
51. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
52. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
53. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. 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)*.
54. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
55. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
56. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
57. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. 10.1007/s00530-010-0182-0.
58. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. http://dx.doi.org/10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
59. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
60. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
61. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
62. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. http://dx.doi.org/10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
63. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
64. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
65. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

 doi:10.20944/preprints202502.1568.v1

66. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
67. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
68. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
69. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
70. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
71. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
72. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
73. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
74. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
75. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
76. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
77. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
78. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
79. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
80. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
81. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
82. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
83. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
84. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
85. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
86. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

87. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
88. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
89. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
90. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
91. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
92. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
93. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
94. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
95. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
96. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.
97. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
98. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
99. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
100. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
101. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
102. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
103. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
104. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
105. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
106. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.

107. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
108. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
109. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
110. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
111. Yu Zhao, Hao Fei, Wei Ji, Jianguo Wei, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Generating visual spatial description via holistic 3D scene understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7960–7977, 2023.