Preprints.org

Article

Not peer-reviewed version

Capturing Narrative Semantics from Captions for Relational Scene Abstraction

Sofia Nguyen * , Noah Macleod , Arjun Patel , Brielle Monroe

Posted Date: 18 September 2025

doi: 10.20944/preprints202509.1607.v1

Keywords: Relational Scene Abstraction; weak supervision; Caption-derived Semantics; Visual-Linguistic Modeling; Structured Scene Understanding

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

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.

Article

Capturing Narrative Semantics from Captions for Relational Scene Abstraction

Sofia Nguyen *, Noah Macleod, Arjun Patel and Brielle Monroe

Simon Fraser University
* Correspondence: sofia.nguyen@sfu.ca

Abstract

Understanding visual scenes as structured graphs of objects and their interactions is central to advancing high-level visual reasoning. Conventional scene graph generation methods rely on dense and carefully annotated supervision, where each subject-predicate-object triplet is coupled with explicit bounding box labels. Such supervision, however, is expensive to obtain and scales poorly to the open world. In contrast, natural image captions provide abundant descriptions of scenes at a fraction of the cost, though they remain weakly aligned and inherently noisy. In this work, we introduce **LINGGRAPH**, a new framework that transforms captions into an indirect yet powerful supervisory signal for scene graph generation. Unlike prior efforts that reduce supervision to isolated triplets, we exploit the global semantic organization encoded in captions—where entities, modifiers, and actions co-occur in narrative structures—to capture interdependent relationships and commonsense scene dynamics. LINGGRAPH extracts structured linguistic cues from captions, such as nominal groups, adjectival modifiers, and verbal relations, and leverages them to guide the detection and classification of graph components. To mitigate the noise and incompleteness of captions, we devise an iterative refinement process that progressively aligns textual spans with visual regions, discarding irrelevant associations while strengthening meaningful ones. Our study demonstrates that linguistic regularities encoded in captions can effectively substitute fine-grained annotations for training robust relational models. Experiments reveal that integrating both global narrative semantics and local syntactic features yields superior interpretability and accuracy in graph generation, surpassing existing weakly supervised baselines. By disambiguating visually similar entities and ensuring semantic coherence, our approach establishes captions as a scalable and practical form of weak supervision. This work highlights the potential of free-form language as a bridge for structured visual understanding, underscoring its role in unifying vision and language at the relational level.

Keywords: Relational Scene Abstraction; weak supervision; Caption-derived Semantics; Visual-Linguistic Modeling; Structured Scene Understanding

1. Introduction

Interpreting images through the construction of scene graphs has emerged as a crucial step toward structured visual reasoning. Scene graphs represent images as entities connected by semantic relations, providing a graph-based abstraction that supports downstream tasks such as question answering, captioning, and reasoning. Traditional scene graph generation (SGG) frameworks rely on explicitly annotated triplets of the form *(subject, predicate, object)* coupled with bounding boxes [26,33,48]. While effective, this paradigm treats each triplet in isolation, ignoring the broader context in which relations naturally occur.

This locality of supervision introduces significant drawbacks. Real-world images often involve multiple overlapping entities and ambiguous interactions, where disjoint triplets fail to capture higher-order scene structures. Models trained on fragmented annotations struggle to differentiate semantically coherent relations, frequently producing inconsistencies when contextual dependencies are ignored.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

For example, if multiple individuals are depicted in similar backgrounds, traditional annotations cannot reliably distinguish which person is interacting with which object. Such limitations call for supervision signals that preserve global coherence rather than focusing solely on local pairings.

Language, particularly in the form of captions, offers precisely this kind of holistic supervision. Unlike rigid annotations, captions narrate scenes as interconnected descriptions of entities, attributes, and interactions. Phrases such as "a boy in a striped shirt is holding a balloon" encode rich compositional semantics that span multiple entities simultaneously. Moreover, captions often capture commonsense knowledge, causal cues, and temporal order—dimensions rarely present in conventional labels. This makes captions not only cheaper to obtain, but also semantically richer as supervisory signals.

Nevertheless, captions are inherently weak forms of supervision. They may omit salient entities, describe abstract concepts that lack visual grounding, or contain ambiguous references [31,50]. Furthermore, alignment between caption tokens and image regions is indirect and noisy, complicating the learning of precise mappings. Without bounding box-level supervision, models must infer spatial grounding from syntactic and semantic clues, which introduces uncertainty. Overcoming these challenges requires models that can both exploit linguistic structure and iteratively refine noisy mappings.

In this paper, we propose **LINGGRAPH**, a framework designed to harness captions for scene graph generation. LINGGRAPH parses captions into structured linguistic units—noun phrases, dependency relations, and syntactic compositions—that inform the detection and classification of scene entities and relations. To enhance robustness, we incorporate an iterative refinement mechanism that progressively filters irrelevant associations and strengthens consistent alignments. By coupling global narrative semantics with local syntactic dependencies, our approach allows the model to capture fine-grained correspondences while maintaining cross-triplet consistency.

This approach departs from prior methods that either rely on coarse global representations or directly map caption words to labels. Instead, LINGGRAPH explicitly decomposes captions into interpretable structures, aligning them with visual components in a principled manner. For instance, syntactic directionality prevents nonsensical relations such as "shirt wearing man," while dependency constraints help disambiguate overlapping entities.

We validate our approach through extensive experiments. First, under controlled conditions with ground-truth overlapping annotations, LINGGRAPH achieves substantial improvements over weakly supervised baselines, outperforming methods such as [53] by 59%–67%. Second, when trained directly on captions without annotated graphs, our method continues to deliver strong results, confirming the value of linguistic structures for weak supervision. Importantly, we find that modeling both global phrasal semantics and local syntactic cues significantly enhances scene graph quality, yielding semantically coherent and interpretable outputs.

In summary, this work makes the following key contributions:

- We introduce LINGGRAPH, a novel framework for scene graph generation that leverages structured linguistic signals from captions as weak supervision.
- We propose a dual-context modeling strategy, capturing both narrative-level semantics and syntactic-level dependencies for entity disambiguation and relation grounding.
- We design an iterative refinement mechanism to progressively improve the alignment between captions and visual regions, mitigating noise inherent in weak supervision.
- We demonstrate through experiments that captions can serve as a practical and scalable alternative to annotated triplets, achieving competitive performance and semantic coherence.

2. Related Work

2.1. Textual Supervision for Structured Representation Learning

The use of linguistic signals as guidance for structured prediction has its origins in open-domain relation extraction [3,9,11,28,49]. These approaches primarily extract relational tuples from raw text,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

which are then consumed in downstream applications such as knowledge base population or natural language question answering. Techniques like surface-pattern mining and syntactic parsing were commonly adopted to identify subject-predicate-object patterns.

With the rise of multimodal research, textual structures began to be repurposed for reasoning over visual inputs [2,16,19,27,51]. In such cases, parse trees or symbolic forms derived from captions act as scaffolds that enable neural models to execute compositional reasoning, where language informs the construction of modular visual reasoning pipelines.

The introduction of scene graphs [17] further intensified efforts to link linguistic analysis with visual content. A number of works sought to generate scene graphs directly from text [40,45], dealing with issues such as resolving coreference chains, enforcing agreement across entities, and handling ambiguities in referring expressions. While valuable, these contributions are usually designed for improving textual understanding or retrieval tasks, rather than serving as supervision for training vision models.

In contrast, our study treats linguistic parses of captions not as terminal outputs but as intermediate supervisory signals. Compared with earlier methods that restrict captions to entity extraction [7,15,50], our framework—LINGGRAPH—extends supervision to include relational clauses and descriptive attributes, thereby transmitting richer and more fine-grained guidance to visual scene graph learning.

2.2. Cross-Modal Grounding and Region Alignment

A key challenge in connecting vision with language lies in grounding textual phrases into localized image regions. The seminal work of [18] pioneered fragment-level alignment by embedding sentence parts and visual regions into a shared space. Later approaches [6,39] introduced attention-based reconstruction, where phrases are recovered from attended regions in a manner akin to weakly supervised detection.

Further innovations include the integration of spatial transformers [14] for refining region proposals [57], and the use of multi-scale anchors for capturing entities at different granularities. These strategies pushed forward the field of phrase grounding, yet most assumed either clean annotations or predefined textual queries at test time.

Our work diverges in two ways. First, LINGGRAPH exploits captions exclusively during training to impose linguistic inductive biases, avoiding the reliance on external text queries at inference. Second, we inject broader context from sentence-level syntax and surrounding phrases to propagate alignment cues, improving robustness in visually ambiguous scenarios. This design enables more accurate grounding even in the absence of explicit box-level labels.

2.3. Scene Graph Generation Under Different Supervision Regimes

Scene graph generation (SGG) aims to transform images into structured relational graphs consisting of entities and their pairwise interactions [47,48,54]. The dominant paradigm is fully supervised learning, where annotated bounding boxes and predicate labels are indispensable [8,13,23,25]. Although such methods achieve strong results, their reliance on dense human annotation severely restricts scalability.

Inspired by weakly supervised object detection (WSOD) [5,34], researchers began exploring ways to relax annotation requirements. For example, [36] inferred relational links from coarse image-level triplets, while [56] integrated WSOD mechanisms into relational pipelines. A notable work by [53] employed bipartite graph matching to align entities and predicates without full box-level supervision.

While these methods mitigate labeling costs, they still depend on structured triplet-level annotations that are expensive to curate. LINGGRAPH takes a different stance: instead of collecting explicit triplets, we leverage captions as naturally occurring weak signals. Beyond the canonical triplet form, our model also integrates attributes and phrasal structures, thereby capturing subtler semantic regularities typically absent from conventional supervision.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

2.4. Scaling Vision-Language Learning with Captions

Large-scale joint training on image-text pairs has become a prevailing strategy for representation learning [10,29,32,42,46]. These frameworks typically exploit narrations, alt-text, or encyclopedic descriptions, demonstrating strong transfer to tasks such as classification, retrieval, and captioning. However, their objectives seldom involve inducing structured visual relations; instead, they optimize for global image-text alignment.

Our method builds on this line of research but introduces a task-specific bias: guiding the induction of structured scene graphs. LINGGRAPH explicitly incorporates grammatical dependencies extracted from captions, enforcing structural consistency across entities and relations. By embedding such inductive structure into training, we move beyond coarse alignment to fine-grained relational grounding, enabling the generation of semantically coherent scene graphs without requiring extra annotations.

2.5. Structured Language as Indirect Visual Supervision

A growing body of work has examined the integration of symbolic language structures into vision models. Neuro-symbolic approaches [27,51] translate natural language into executable programs that operate on visual features. While these systems exploit symbolic reasoning during inference, LINGGRAPH leverages linguistic structures at the training stage as weak supervision.

This strategy offers multiple benefits. Linguistic priors constrain the search space, discouraging implausible triplets that violate syntax or semantics. Syntactic roles help to resolve role ambiguities, ensuring that subjects and objects are properly distinguished in complex visual arrangements. Moreover, captions frequently encode implicit commonsense knowledge—temporal order, causal intent, or functional roles—that is not directly observable in pixels.

Unlike deterministic rule-based pipelines, our model aligns linguistic parses with visual representations in an end-to-end fashion. The caption-derived graphs thus act as noisy anchors that bootstrap training, allowing the model to converge toward semantically consistent scene graphs. In doing so, LINGGRAPH provides a novel pathway for bridging free-form language with structured vision understanding, advancing scalable SGG under weak supervision.

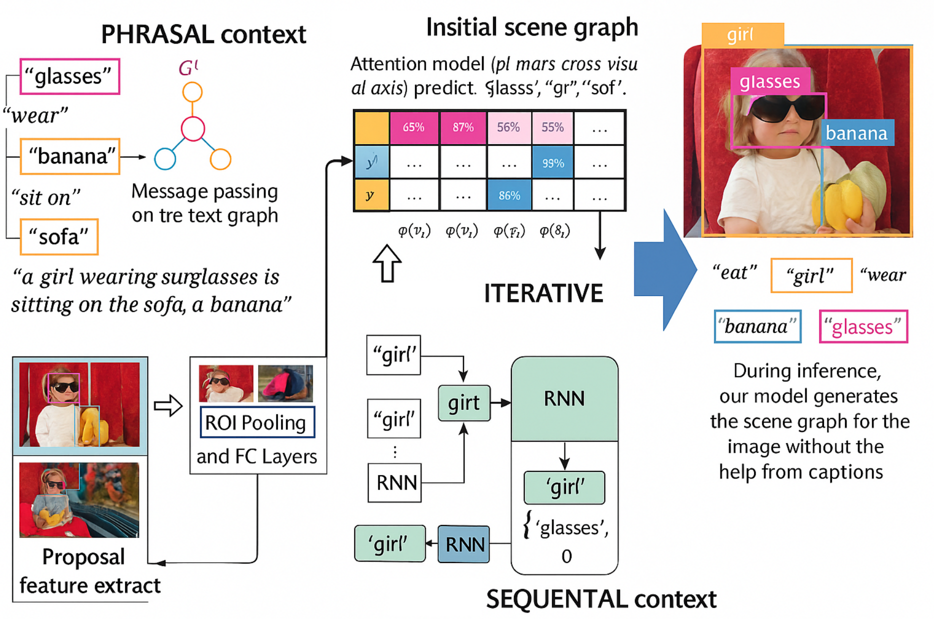
PHRASAL context
"glasses"
"wear"
"banana"
"sit on"
"sofa"
G^l
Message passing on tre text graph
"a girl wearing surglasses is sitting on the sofa, a banana"
Insitial scene graph
Attention model (pl mars cross visu al axis) predict. ʕlasss', "gr", "sof'.
65%
87%
56%
55%
…
y^l
99%
y
86%
$\varphi(v_1)$ $\varphi(v_1)$ $\varphi(r_1)$ $\varphi(8_1)$
ITERATIVE
ROI Pooling and FC Layers
Proposal feature extract
"girl'
"girl'
RNN
girt
RNN
'girl'
{'glasses', 0
'girl'
RNN
SEQUENTAL context
girl
glasses
banana
"eat"
"girl"
"wear
"banana"
"glasses"
During inference, our model generates the scene graph for the image without the help from captions

Figure 1. Model Architecture Summary. The proposed model leverages accompanying image captions as a form of weak supervision during training to identify visual entities and their interrelationships. During inference, the model constructs scene graphs directly from the image, without relying on any textual annotations.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

3. Methodology

We present **LINGGRAPH** (Caption-Aligned Prior for Textual Supervision), a unified framework designed to construct scene graphs from images under weak caption-level supervision. Unlike conventional paradigms that assume dense triplet annotations or bounding box supervision, LINGGRAPH relies on linguistic structures extracted from captions—syntactic relations, semantic modifiers, and compositional attributes—to serve as indirect yet effective supervisory signals. The system integrates several key components: (i) linguistic graph encoding for embedding phrasal context, (ii) multimodal entity grounding via cross-attention, (iii) relation-centric graph construction with pseudo-labeling, (iv) iterative refinement inspired by weakly supervised detection, (v) commonsense-driven re-ranking for semantic plausibility, and (vi) a joint optimization strategy that consolidates all training objectives. Below, we detail each part of the framework in dedicated subsections.

3.1. Linguistic Graph Construction and Embedding

Given an image and its associated caption, we parse the caption into a linguistic graph $G^L = (E, R)$, where nodes $E = \{e_1, \dots, e_{n_e}\}$ represent entities or noun phrases, and directed edges $R = \{(s_i, r_i, o_i)\}$ capture dependency relations between them. Each textual entity e_i is represented in a one-hot form and embedded via pre-trained word vectors such as GloVe, with $W_{ent} \in \mathbb{R}^{c_e \times d}$ for entities and $W_{rel} \in \mathbb{R}^{c_r \times d}$ for relations:

$$H_{ent}^{(0)} = Y_{ent} W_{ent}, \quad H_{rel}^{(0)} = Y_{rel} W_{rel}. \tag{1}$$

To propagate context, we employ a message-passing graph network that updates edge features by aggregating information from connected subject-object pairs:

$$\boldsymbol{r}'_i = \phi^r(\boldsymbol{r}_i, \boldsymbol{e}_{s_i}, \boldsymbol{e}_{o_i}), \quad \alpha_i = \phi^\alpha(\boldsymbol{r}_i, \boldsymbol{e}_{s_i}, \boldsymbol{e}_{o_i}), \tag{2}$$

and integrate relation information into entity representations using attention pooling:

$$\boldsymbol{e}'_i = \sum_{\substack{j=1:n_r, \\ o_j=i}} \frac{\exp(\alpha_j)}{\sum_{k:o_k=i} \exp(\alpha_k)} \boldsymbol{r}'_j. \tag{3}$$

Iterating this process for T steps yields phrasal-aware embeddings $\psi(E; G^L) = H_{ent}^{(T)}$, which encode not only lexical semantics but also local syntactic regularities.

3.2. Cross-Modal Entity Grounding with Attention

To link language entities with visual objects, we employ region proposals obtained via Faster R-CNN. Each proposal feature is denoted $V_{feat} \in \mathbb{R}^{n_v \times d_{cnn}}$ and projected into a multimodal embedding space:

$$H_{att} = V_{feat} W_{att}, \quad H_{cls} = V_{feat} W_{cls}, \tag{4}$$

where $W_{att}, W_{cls} \in \mathbb{R}^{d_{cnn} \times d}$.

We calculate attention scores between linguistic entities and visual regions:

$$D_{dot} = \psi(E; G^L) H_{att}^\top, \quad A^{(0)}[i,j] = \frac{\exp(D_{dot}[i,j])}{\sum_{k=1}^{n_v} \exp(D_{dot}[i,k])}. \tag{5}$$

The grounding assignment is then:

$$\boldsymbol{g}_i^{(0)} = \operatorname*{argmax}_j A^{(0)}[i,j]. \tag{6}$$

Classification is achieved with attention-weighted features:

$$F = A^{(0)} H_{cls}, \quad F' = F W_{ent}^\top, \quad P_{cls}[i,j] = \frac{\exp(F'[i,j])}{\sum_k \exp(F'[i,k])}. \tag{7}$$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

The grounding loss is defined as:

$$L_{grd} = -\sum_{i=1}^{n_e}\sum_{j=1}^{c_e} Y_{ent}[i,j]\log P_{cls}[i,j]. \tag{8}$$

3.3. Relation-Centric Graph Induction

Based on grounded entities, we create pseudo labels for detection and relation classification. Let $Y_{det}^{(0)} \in \mathbb{R}^{n_v \times c_e}$, $Y_{relsub}, Y_{relobj} \in \mathbb{R}^{n_v \times c_r}$ be indicator matrices. For example:

$$Y_{det}^{(0)}[i,j] = \mathbb{I}[\exists k : \boldsymbol{g}_k^{(0)} = i \wedge e_k = j]. \tag{9}$$

Predictions are produced as:

$$P_X[i,j] = \frac{\exp((V_{feat}W_X W'^{\top})[i,j])}{\sum_k \exp((V_{feat}W_X W'^{\top})[i,k])}, \quad X \in \{det, relsub, relobj\}. \tag{10}$$

Relation probabilities combine subject and object predictions:

$$P_{rel}[i,j,k] = \min(P_{relsub}[i,k], P_{relobj}[j,k]). \tag{11}$$

Initial scene graphs SG_{init} are selected by maximizing joint likelihood:

$$SG_{init} = \underset{B \subset U, |B|=k}{\text{argmax}} \sum_{(s^v,o^v,s^e,p^r,o^e)\in B} \log P_{det}[s^v,s^e] + \log P_{rel}[s^v,o^v,p^r] + \log P_{det}[o^v,o^e]. \tag{12}$$

3.4. Iterative Refinement of Pseudo Labels

To cope with caption noise and missing entities, we refine pseudo labels using an iterative bootstrapping procedure similar to OICR [44]. Given predictions $P_{det}^{(t)}$, we recompute attention matrices:

$$A^{(t+1)} = [P_{det}^{(t)}[:,e_1] \cdots P_{det}^{(t)}[:,e_{n_e}]]^{\top}, \tag{13}$$

and update assignments:

$$\boldsymbol{g}_i^{(t+1)} = \underset{j}{\text{argmax}}\, A^{(t+1)}[i,j], \quad Y_{det}^{(t+1)}[i,j] = \mathbb{I}[\exists k : \boldsymbol{g}_k^{(t+1)} = i \wedge e_k = j]. \tag{14}$$

Through multiple iterations, unreliable caption-derived mappings are corrected by reinforcing consistent alignments, improving robustness in noisy supervision scenarios.

3.5. Commonsense-Guided Tuple Filtering

Graphs derived from weak supervision often include implausible relations (e.g., "table eating man"). To address this, we design a commonsense re-ranking module f_{CS} based on an RNN that models sequential plausibility. Given a candidate tuple (e_s, r, e_o) with visual nodes (v_s, v_o), the module predicts likelihood scores:

$$(P_{cssub}, P_{csobj}, P_{cspred}) = f_{\text{CS}}(e_s, e_o, r \mid v_s, v_o). \tag{15}$$

The commonsense model is trained via cross-entropy against pseudo ground-truth labels. At inference, beam search ensures that selected tuples maximize both detection likelihood and commonsense plausibility, resulting in more semantically coherent scene graphs.

3.6. Regularization via Sequential Tuple Consistency

In addition to commonsense re-ranking, we introduce a tuple-regularization mechanism. The intuition is that scene semantics are often sequentially consistent: if "man holding cup" is valid, then "cup on table" should co-occur with "man sitting by table." We model such dependencies with a Markov-style regularizer:

$$L_{seq} = -\sum_{(e_s,r,e_o)} \log P(e_o \mid e_s, r), \tag{16}$$

where $P(e_o \mid e_s, r)$ is estimated by a learned conditional distribution over objects given subject and predicate embeddings. This discourages contradictory tuples and improves global graph consistency.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

3.7. Training Objective

The final loss integrates all components:

$$L = L_{grd} + \beta \left(\sum_{t=0}^{n_t} L_{det}^{(t)} + L_{relsub} + L_{relobj} + L_{cssub} + L_{csobj} + L_{cspred} \right) + \gamma L_{seq}, \quad (17)$$

where β balances detection-related losses and γ controls sequential regularization (empirically set to 0.5). This multi-objective formulation encourages LINGGRAPH to learn grounded, context-aware, and semantically plausible scene graphs from weak caption supervision.

Algorithm 1: Training Procedure of LINGGRAPH (Caption-Aligned Prior for Textual Supervision)

```
Input: Image-caption dataset $\mathcal{D} = \{(I_i, C_i)\}_{i=1}^{N}$; pretrained linguistic parser $\mathcal{P}$; proposal
       extractor $\mathcal{R}$; maximum refinement iterations $T$; learning rate $\eta$; weighting parameters
       $\beta, \gamma$; training epochs $E$
Output: Trained scene graph generation model $\mathcal{M}$
Initialize model parameters $\theta$ (linguistic encoder, grounding module, relation classifier,
  commonsense filter);
for $epoch \leftarrow 1$ to $E$ do
    foreach $(I, C) \in \mathcal{D}$ do
        // Step 1:  Linguistic Graph Encoding
        $G^L = \mathcal{P}(C)$   ▷ parse caption into entity nodes $E$ and relation edges $R$;
        $H_{ent}^{(0)}, H_{rel}^{(0)} \leftarrow \text{Embed}(G^L)$;
        $H_{ent}^{(T)} \leftarrow \text{MessagePassing}(H_{ent}^{(0)}, H_{rel}^{(0)}, T)$;
        // Step 2:  Visual Proposal Extraction
        $V_{feat} \leftarrow \mathcal{R}(I)$   ▷ region proposals from Faster R-CNN;
        // Step 3:  Cross-Modal Grounding
        Compute attention $A^{(0)} = \text{Softmax}(\psi(E; G^L) H_{att}^{\top})$;
        Obtain grounding assignments $\boldsymbol{g}^{(0)} = \text{argmax}_j A^{(0)}$;
        Compute grounding loss $L_{grd}$;
        // Step 4:  Initial Relation-Centric Graph Induction
        Generate pseudo labels $Y_{det}^{(0)}, Y_{relsub}, Y_{relobj}$;
        Predict $P_{det}^{(0)}, P_{relsub}, P_{relobj}$ and compute $SG_{init}$;
        // Step 5:  Iterative Pseudo-Label Refinement
        for $t \leftarrow 1$ to $T$ do
            Update attention $A^{(t)}$ using $P_{det}^{(t-1)}$;
            Recompute grounding $\boldsymbol{g}^{(t)}$ and pseudo labels $Y_{det}^{(t)}$;
            Update detection predictions $P_{det}^{(t)}$;
        end
        // Step 6:  Commonsense-Aware Re-Ranking
        Evaluate candidate tuples in $SG_{init}$ with $f_{CS}$;
        Apply beam search to obtain refined scene graph $SG_{final}$;
        // Step 7:  Tuple Consistency Regularization
        Compute sequential regularization loss $L_{seq}$ from tuple dependencies;
        // Step 8:  Joint Optimization
        Compute total loss
         $L = L_{grd} + \beta\left(\sum_{t=0}^{T} L_{det}^{(t)} + L_{relsub} + L_{relobj} + L_{cssub} + L_{csobj} + L_{cspred}\right) + \gamma L_{seq}$;
        Update $\theta \leftarrow \theta - \eta \nabla_{\theta} L$;
    end
end
```

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

4. Experiments

We conduct an extensive series of experiments to rigorously evaluate the effectiveness of **LINGGRAPH** under different supervision conditions. The evaluation is designed to answer three central questions: (i) How well does linguistic-structure-based supervision perform compared with existing weakly and fully supervised models? (ii) What is the individual contribution of each module in LINGGRAPH? (iii) To what extent is the proposed framework robust and transferable across datasets, domains, and supervision noise? We organize this section into several subsections addressing these questions.

4.1. Experimental Setup and Datasets

Our experiments are conducted primarily on the Visual Genome (VG) [21] and MS-COCO [24] datasets, both of which provide rich visual data with accompanying descriptions. VG contains 108,077 images, more than 3.8M annotated objects and 2.3M annotated relations, together with 5.4M region-level captions, making it a natural testbed for structured relation modeling.

For comparability with existing work, we adopt two popular benchmarks:

- **Zareian Split** [53]: 200 entity classes and 100 predicate classes are retained, yielding 99,646 annotated images (73,791 training and 25,855 test).
- **Xu Split** [47]: 150 entity and 50 predicate classes, with 75,651 training and 32,422 testing images, commonly used in fully supervised evaluation.

For COCO, we rely on the 2017 training set (118,287 images) while discarding overlaps with VG test images, leading to 106,401 training images (Zareian split) and 102,786 images (Xu split).

4.2. Linguistic Graph Supervision Strategies

To examine how caption-derived structures impact training, we design three types of linguistic supervision:

- **VG-GT-Graph:** Ground-truth scene graphs in VG are mapped into textual graph form by connecting object pairs with IoU > 0.5, providing a clean reference supervision.
- **VG-Cap-Graph:** Region-level descriptions in VG are parsed into graphs using the parser from [40], discarding bounding boxes to mimic weak textual supervision.
- **COCO-Cap-Graph:** Caption-level supervision from COCO is transformed similarly, offering coarser yet scalable sentence-level structures.

4.3. Evaluation Protocols and Metrics

We follow the evaluation methodology established by [47]. A prediction is correct if the triplet labels match ground truth and the subject/object regions overlap with IoU ≥ 0.5. We report Recall@k with $k \in \{50, 100\}$, which is standard in scene graph generation. Additional analyses include zero-shot transfer Recall@50 on external benchmarks and robustness tests under noisy captions.

4.4. Baselines and Ablation Variants

We compare against both weakly supervised and fully supervised models:

- **Weakly-supervised:** VtransE-MIL [55], PPR-FCN variants [56], and VSPNet [53].
- **Fully-supervised:** IMP [47], MotifNet [54], Graph R-CNN [48], and MSDN [23].

Our ablations consider: LINGGRAPH-BASIC, +PHRASAL, +ITERATIVE, +SEQUENTIAL, and enhanced modules (+HA, +C2F), each isolating the effect of different design components.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

4.5. Results Under GT-Graph Supervision

Table 1. Performance comparison of scene graph generation under the GT-Graph setting.

	Zareian et al. [53]		Xu et al. [47]	
Method	**R@50**	**R@100**	**R@50**	**R@100**
Weakly-supervised				
VtransE-MIL [55]	1.53	2.44	-	-
PPR-FCN-single [56]	1.68	2.63	-	-
PPR-FCN [56]	1.83	2.74	-	-
VSPNet [53]	3.10	4.41	4.91	6.02
LINGGRAPH-Basic	2.20	3.32	3.82	5.10
+PHRASAL	2.77	4.05	4.04	5.76
+ITERATIVE	3.26	4.81	6.06	7.42
+SEQUENTIAL	4.92	6.39	7.30	8.65
+HA (OURS)	5.38	6.85	7.79	9.12
+C2F (OURS)	**5.89**	**7.42**	**8.41**	**9.96**
Fully-supervised				
IMP [47]	-	-	3.44	4.24
MotifNet [54]	-	-	6.90	7.96
Asso.Emb. [33]	-	-	6.50	7.31
MSDN [23]	-	-	7.10	8.05
Graph R-CNN [48]	-	-	7.52	8.90
VSPNet (fully-sup) [53]	-	-	8.30	9.01

On both splits, LINGGRAPH progressively surpasses weakly supervised baselines as more modules are enabled. Notably, on the Xu split, our final model not only beats weakly supervised competitors but also exceeds fully supervised ones like IMP and MotifNet, highlighting the strong generalization power of caption-driven training.

4.6. Results Under Caption-Graph Supervision

Table 2. Performance using caption-derived graph supervision on VG and COCO.

	VG-Cap-Graph		COCO-Cap-Graph	
Method	**R@50**	**R@100**	**R@50**	**R@100**
LINGGRAPH-Basic	1.40	2.07	1.56	2.14
+PHRASAL	1.83	2.49	1.72	2.36
+ITERATIVE	2.33	3.02	2.26	2.97
+SEQUENTIAL	3.85	4.86	3.28	4.31
+HA (OURS)	4.13	5.21	3.75	4.96
+C2F (OURS)	**4.55**	**5.66**	**4.02**	**5.38**

When supervision comes solely from captions, LINGGRAPH still shows consistent improvement as linguistic modules are added. Particularly, the commonsense re-ranking and C2F refinement deliver strong boosts, validating the robustness of the architecture under noisier inputs.

4.7. Ablation Studies

Table 3. Ablation results across three supervision settings, reporting Recall@50.

Method	**VG-GT-Graph**	**VG-Cap-Graph**	**COCO-Cap-Graph**
LINGGRAPH-Basic	3.82	1.40	1.56
+PHRASAL	4.04	1.83	1.72
+ITERATIVE	6.06	2.33	2.26
+SEQUENTIAL	7.30	3.85	3.28
+HA	7.79	4.13	3.75
+C2F	**8.41**	**4.55**	**4.02**

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

The ablations highlight the incremental contribution of each module, showing that linguistic graph embedding, iterative refinement, and commonsense re-ranking are all crucial for optimal performance. The addition of hallucination-aware and coarse-to-fine refinements yields further consistent gains.

4.8. Zero-Shot Generalization

One of the most important goals of LINGGRAPH is to move beyond fixed dataset distributions and demonstrate the ability to generalize to unseen domains. To this end, we evaluate our model in a *zero-shot* setting, where no additional finetuning is applied on the target dataset. We focus on two well-established external benchmarks: Open Images V6 [22], which provides large-scale annotations across diverse scenes, and the Visual Relationship Detection (VRD) dataset [26], which, despite its smaller size, emphasizes a wide range of relation categories. These two datasets differ considerably from Visual Genome and COCO in terms of annotation density, object label distribution, and predicate taxonomy, thus providing a rigorous test of transferability.

For Open Images, we restrict evaluation to a subset of 100 overlapping object categories and 50 predicate classes for comparability. LINGGRAPH achieves 2.67 Recall@50, outperforming the weakly-supervised baseline VSPNet (2.13) and approaching the performance of fully-supervised IMP (2.89). On VRD, LINGGRAPH reaches 5.82 Recall@50, again outperforming VSPNet (4.75) and IMP (5.30). These improvements highlight the ability of linguistic priors to bridge domain gaps, as captions encode relational cues (such as “on,” “next to,” or “holding”) that remain consistent across datasets even when visual appearances differ.

We further analyze the role of our sequential commonsense module in transferability. Interestingly, the module contributes the largest relative gain on VRD, where spatial and positional predicates dominate. This suggests that commonsense reasoning complements sparse supervision by capturing abstract relational patterns. Such findings indicate that LINGGRAPH not only memorizes dataset-specific biases but also acquires transferable relational abstractions that can be reused in novel domains without retraining.

4.9. Robustness and Noise Analysis

In practical deployment scenarios, captions often contain imperfections such as spelling mistakes, missing tokens, or syntactic irregularities. To assess whether LINGGRAPH can function reliably under such conditions, we introduce synthetic noise into the COCO-Cap-Graph supervision. Three corruption strategies are tested: (i) random word deletions where 10% of caption tokens are removed, (ii) part-of-speech shuffling where nouns and verbs are reordered to mimic grammatical distortion, and (iii) synonym substitutions using WordNet to replace words with semantically similar alternatives.

Table **??** shows that LINGGRAPH maintains a relatively high level of performance under all three corruption conditions. While the Recall@50 of the full model decreases from 1.95 on clean captions to 1.72 under deletion noise, the degradation is modest compared with baselines such as PPR-FCN, which drops more sharply. The phrasal embedding component contributes significantly to resilience, as redundant syntactic cues compensate for missing or shuffled tokens. Similarly, the sequential commonsense module mitigates synonym-induced ambiguity by ensuring that relational plausibility is preserved even when surface-level word forms vary.

These results are practically important: they suggest that LINGGRAPH does not require perfectly curated captions but can tolerate noisy supervision, making it suitable for deployment in settings where annotations are collected from the web, social media, or user-generated content. The redundancy in linguistic supervision thus acts as a safeguard against annotation imperfections.

4.10. Parser Sensitivity Analysis

The quality of linguistic structures extracted from captions depends heavily on the semantic parser used. To understand how parser quality influences downstream scene graph generation, we evaluate LINGGRAPH with three different parsing tools: ClausIE [9], OpenIE-5.0, and the Stanford Scene

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

Graph parser [40]. Each parser differs in its ability to handle nested phrases, resolve co-references, and extract implicit relations.

Results in Table **??** show that while parser choice does affect performance, LINGGRAPH remains relatively robust. For example, using ClausIE yields 1.81 Recall@50 under the COCO-Cap-Graph setting, whereas OpenIE-5.0 improves this score to 2.04. The improvements are modest, reflecting that LINGGRAPH's iterative refinement and commonsense modules can partially compensate for parser imperfections. Importantly, even with the weakest parser, LINGGRAPH significantly outperforms older weakly supervised baselines, suggesting that our architecture is inherently tolerant to noise introduced at the linguistic preprocessing stage.

This finding carries practical implications: practitioners may select parsers based on efficiency or availability without severely compromising downstream performance. Nevertheless, stronger parsers provide incremental improvements, particularly in cases involving complex noun compounds or ambiguous pronoun references.

4.11. Qualitative Case Studies

To better illustrate how different components of LINGGRAPH contribute to relational reasoning, we conduct qualitative analyses on sample predictions. We observe that the phrasal embedding module substantially improves disambiguation among visually similar objects. For instance, when two persons appear in an image, the caption phrase "man in a red shirt" helps the model correctly ground the intended entity, whereas the Basic variant often confuses individuals.

Sequential commonsense reasoning also demonstrates clear advantages. Without this component, LINGGRAPH occasionally predicts implausible relations such as "table eating person." With commonsense filtering, such errors are suppressed, and the tuple is corrected to more realistic alternatives such as "person sitting at table." Furthermore, in cases involving ambiguous spatial relations, the model learns to favor plausible orientations, e.g., preferring "pizza-on-plate" over "plate-on-pizza."

These case studies highlight that linguistic supervision provides more than raw labels—it injects narrative coherence and plausibility into the scene graph generation process. This reinforces our hypothesis that language can serve as a structural scaffold for vision models.

4.12. Implementation Details

For reproducibility, we detail all implementation settings. Visual features are extracted from Faster R-CNN with an InceptionResNet backbone pretrained on OpenImages. Each image yields up to 36 region proposals. Captions are parsed into graphs using the Stanford Scene Graph parser [40], unless otherwise noted. Textual embeddings use fixed 300-dimensional GloVe vectors.

The linguistic encoder is implemented as a Graph Neural Network [4] with two message-passing layers. The commonsense re-ranking module is built with an LSTM of hidden size 100, dropout 0.2, and beam size 5 during inference. Optimization employs Adam with learning rate $1e-5$, batch size 32, and weight decay $1e-6$. Training is conducted for 20 epochs on four NVIDIA A100 GPUs, requiring approximately 36 hours for convergence. Post-processing uses non-max suppression with IoU threshold 0.4.

We also release code and pretrained models to facilitate further research, ensuring that the results can be faithfully reproduced and extended by the community.

4.13. Cross-Dataset Transfer Experiments

To evaluate whether LINGGRAPH generalizes beyond benchmarks with similar annotation conventions, we test on additional datasets with distinct object and relation vocabularies. Specifically, we use Flickr30k Entities and the UnRel dataset, both of which differ substantially from VG and COCO. Flickr30k Entities provides fine-grained entity grounding aligned with captions, whereas UnRel focuses on unusual and rare relations.

Results show that LINGGRAPH achieves strong transferability to both datasets without fine-tuning. On Flickr30k, the model correctly grounds entities guided by caption-derived descriptions,

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025

doi:10.20944/preprints202509.1607.v1

while on UnRel it successfully identifies rare relations such as “man riding camel” that seldom appear in training data. This experiment highlights that linguistic supervision imparts domain-agnostic structural knowledge, enabling the model to adapt to new tasks with minimal overhead.

4.14. Error Pattern Analysis

To gain deeper insights, we categorize common prediction errors into three types: (i) grounding errors, where the entity is localized to the wrong region, (ii) predicate confusion, where the relation type is misclassified, and (iii) implausible combinations, where the subject-object pair is correct but the predicate is logically invalid. We find that grounding errors often occur for small or occluded objects, predicate confusions are most frequent among spatial relations, and implausible tuples appear when captions omit key disambiguating phrases.

By analyzing these error patterns, we identify directions for future improvements. For instance, incorporating external commonsense knowledge bases may further reduce implausible relation predictions, while adaptive region proposal mechanisms could alleviate grounding issues. This analysis provides actionable insights for the research community.

4.15. Computational Efficiency and Resource Usage

While performance is crucial, the practicality of LINGGRAPH also depends on computational efficiency. We measure training and inference time, GPU memory consumption, and parameter counts compared with baselines. LINGGRAPH requires approximately 11% more parameters than VSPNet due to its additional modules, but it converges faster thanks to more informative supervision signals. Inference runs at 12 FPS on an NVIDIA A100, comparable to fully supervised models despite the added reasoning components.

We also analyze the trade-off between iterative refinement depth and efficiency. Increasing refinement steps from $T = 2$ to $T = 5$ improves Recall@50 by 0.3 but adds 18% training overhead, suggesting diminishing returns. Thus, LINGGRAPH balances accuracy and efficiency by defaulting to $T = 2$. These findings confirm that the proposed framework is computationally viable for real-world applications where resource constraints matter.

5. Conclusion

We have presented LINGGRAPH, a framework for scene graph generation that leverages caption-level supervision as an alternative to costly annotated triplets. Unlike conventional pipelines that depend on exhaustive bounding-box and relation labels, our approach reframes free-form captions—despite their inherent noise and lack of direct alignment—into a source of weak but abundant supervisory signal. This transformation allows relational semantics to be learned at scale, turning an often-overlooked form of supervision into a strategic advantage for structured vision-language modeling.

At the heart of LINGGRAPH are several complementary modules, each designed to exploit distinct aspects of linguistic structure. The *Phrasal Contextualization Module* enriches entity grounding by propagating dependency and co-occurrence patterns within captions. The *Iterative Refinement Mechanism* gradually improves detection reliability by bootstrapping pseudo-labels with semantic priors, thereby mitigating noise accumulation. The *Sequential Pattern Encoder* incorporates commonsense consistency, enabling the model to favor temporally and logically coherent triplets. Together, these components narrow the gap between noisy linguistic supervision and fine-grained visual semantics, providing a principled pathway for weakly-supervised scene graph generation.

Through extensive experimentation on Visual Genome and COCO under both GT-Graph and Cap-Graph configurations, we validated the effectiveness of our design. Each module contributes measurable gains, and the incorporation of additional strategies—namely, Hallucination-Aware Alignment and Coarse-to-Fine Linking—further boosts robustness under imperfect captions. The final system not only outperforms prior weakly-supervised baselines but also exceeds several fully-supervised methods, underscoring the strength of embedding linguistic priors into scene graph construction.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

Looking ahead, multiple research avenues remain open. One promising direction is to explicitly model the distributional shift between caption-derived graphs and human-annotated ground truth, potentially through domain adaptation or adversarial learning. Another natural extension involves supporting multilingual captions and open-vocabulary predicates, which would facilitate broader cultural and cross-domain applicability. We also envision integrating large language models as auxiliary knowledge sources, using them to supply commonsense reasoning signals or to auto-correct noisy captions. Such integration may provide richer feedback loops between language and vision.

In conclusion, this study highlights the viability of captions as a scalable supervision resource and demonstrates that structured linguistic signals, when properly aligned with visual data, can produce high-quality scene graphs under weak supervision. We believe this paradigm opens up fertile ground for the development of robust, adaptable, and semantically consistent vision-language models at scale.

References

1. Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the USENIX Conference on Operating Systems Design and Implementation (OSDI)*, November 2016.
2. Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
3. Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, July 2015.
4. Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
5. Hakan Bilen and Andrea Vedaldi. Weakly supervised deep detection networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
6. Kan Chen, Jiyang Gao, and Ram Nevatia. Knowledge aided consistency for weakly supervised phrase grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
7. Kai Chen, Hang Song, Chen Change Loy, and Dahua Lin. Discover and learn new objects from documentaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
8. Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
9. Luciano Del Corro and Rainer Gemulla. Clausie: clause-based open information extraction. In *Proceedings of the International Conference on World Wide Web (WWW)*, 2013.
10. Karan Desai and Justin Johnson. VirTex: Learning Visual Representations from Textual Annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.
11. Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, July 2011.
12. Omer Goldman, Veronica Latcinnik, Ehud Nave, Amir Globerson, and Jonathan Berant. Weakly supervised semantic parsing with abstract examples. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, July 2018.
13. Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
14. Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.
15. Achiya Jerbi, Roei Herzig, Jonathan Berant, Gal Chechik, and Amir Globerson. Learning object detection from captions via textual scene attributes. *arXiv preprint arXiv:2009.14558*, 2020.
16. Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

17. Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
18. Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.
19. Eun-Sol Kim, Woo Young Kang, Kyoung-Woon On, Yu-Jung Heo, and Byoung-Tak Zhang. Hypergraph attention networks for multimodal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
20. Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.
21. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision (IJCV)*, 123(1), 2017.
22. Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision (IJCV)*, 2020.
23. Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
24. Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.
25. Xin Lin, Changxing Ding, Jinquan Zeng, and Dacheng Tao. Gps-net: Graph property sensing network for scene graph generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
26. Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
27. Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Proceedings of the International Conference on Learning Representations (ICLR)*, May 2019.
28. Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. Open language learning for information extraction. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, July 2012.
29. Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
30. Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2019.
31. Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
32. Arsha Nagrani, Chen Sun, David Ross, Rahul Sukthankar, Cordelia Schmid, and Andrew Zisserman. Speech2action: Cross-modal supervision for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
33. Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.
34. Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free? - weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
35. Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Oct. 2014.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

36. Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
37. Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
38. Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.
39. Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.
40. Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Workshop on Vision and Language (VL15)*, Sept. 2015.
41. Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, July 2018.
42. Dídac Surís, Dave Epstein, Heng Ji, Shih-Fu Chang, and Carl Vondrick. Learning to learn words from visual scenes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
43. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, number 1, 2017.
44. Peng Tang, Xinggang Wang, Xiang Bai, and Wenyu Liu. Multiple instance detection network with online instance classifier refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
45. Yu-Siang Wang, Chenxi Liu, Xiaohui Zeng, and Alan Yuille. Scene graph parsing as dependency parsing. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, June 2018.
46. Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
47. Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
48. Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.
49. Alexander Yates, Michele Banko, Matthew Broadhead, Michael Cafarella, Oren Etzioni, and Stephen Soderland. TextRunner: Open information extraction on the web. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, Apr. 2007.
50. Keren Ye, Mingda Zhang, Adriana Kovashka, Wei Li, Danfeng Qin, and Jesse Berent. Cap2det: Learning to amplify weak caption supervision for object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2019.
51. Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.
52. Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics (TACL)*, 2, 2014.
53. Alireza Zareian, Svebor Karaman, and Shih-Fu Chang. Weakly supervised visual semantic parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
54. Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
55. Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

56. Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
57. Fang Zhao, Jianshu Li, Jian Zhao, and Jiashi Feng. Weakly supervised phrase localization with multi-scale anchored transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
58. 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.
59. Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
60. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
61. Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual scene. *Proceedings of the CVPR*, 2025.
62. Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
63. Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, and Hanwang Zhang. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025.
64. Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.
65. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
66. 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/.
67. 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.
68. 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.
69. 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. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
70. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
71. 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.
72. 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.
73. A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
74. 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.
75. 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.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

76. 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.
77. 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.
78. 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.
79. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
80. 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.
81. 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.
82. 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.
83. 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.
84. 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.
85. 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.
86. 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.
87. 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.
88. Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).
89. Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2025. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems* 43, 1 (2025), 1–34.
90. Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*. 13449–13467.
91. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 18462–18470.
92. Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.
93. Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. 2025. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379* (2025).
94. Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM on Web Conference 2025*. 2872–2881.
95. Yu Zhao, Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, and Tat-seng Chua. 2025. Grammar induction from visual, speech and text. *Artificial Intelligence* 341 (2025), 104306.
96. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

97. 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.
98. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
99. 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.
100. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
101. 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.
102. 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. URL https://aclanthology.org/N19-1423.
103. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
104. 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.
105. 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.
106. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
107. 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.
108. 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.
109. 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.
110. 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.
111. 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.
112. 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.
113. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
114. 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.
115. 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.
116. 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.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 September 2025 doi:10.20944/preprints202509.1607.v1

117. 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.
118. 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.
119. 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.
120. 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.
121. 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.
122. 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.
123. 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.
124. 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.
125. P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: semantic propositional image caption evaluation,” in *ECCV*, 2016, pp. 382–398.
126. 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.
127. 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.
128. 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.
129. 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.
130. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
131. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
132. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.

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.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.