Preprints.org

Article

Not peer-reviewed version

Collective Semantic Synthesis for Multimodal Alignment through Cognitive Consensus Modeling

Clara Dupont *, Hugo Bernard, Saidi Kareem, Emilie Garnier

Posted Date: 16 October 2025

doi: 10.20944/preprints202510.1295.v1

Keywords: cognitive consensus; multimodal alignment; semantic synthesis; conceptual reasoning; collective knowledge modeling

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Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1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

Collective Semantic Synthesis for Multimodal Alignment through Cognitive Consensus Modeling

Clara Dupont *, Hugo Bernard, Saidi Kareem and Emilie Garnier

Aix-Marseille University
* Correspondence: clara.dupont@univ-amu.fr

Abstract

Establishing a unified representational ground between vision and language remains one of the most persistent challenges in artificial intelligence. Existing image–text alignment paradigms primarily depend on explicit instance-level correspondences derived from paired data, which capture surface associations but neglect the deeper web of conceptual regularities that guide human understanding. In human cognition, interpretation is not a direct reflection of observation but a synthesis of shared experience—a consensus of how objects, actions, and contexts interrelate. The absence of such cognitive consensus in current systems hinders robustness, interpretability, and adaptability across domains. To address this limitation, we propose **SYMBOL** (**S**emantic–s**Y**nthesis via **M**ultimodal **B**ehavioral kn**O**wledge **L**inking), a new framework that aligns visual and linguistic semantics through collective concept integration. SYMBOL constructs a *Cognitive Consensus Graph (CCG)* derived from large-scale multimodal corpora, encoding co-occurrence regularities that emerge from collective human annotations and narrative descriptions. By propagating conceptual relationships across this graph, SYMBOL enriches conventional instance-level representations with consensus-driven knowledge priors. The resulting embedding space jointly models explicit perceptual alignment and implicit conceptual reasoning, enabling more resilient and cognitively coherent multimodal understanding. Comprehensive evaluations on multiple retrieval benchmarks demonstrate that SYMBOL significantly outperforms contemporary systems in bidirectional retrieval and cross-domain adaptation. Beyond quantitative gains, SYMBOL illustrates a new principle: integrating consensus-based semantic synthesis can transform multimodal learning from mere correlation fitting into cognitively grounded reasoning.

Keywords: cognitive consensus; multimodal alignment; semantic synthesis; conceptual reasoning; collective knowledge modeling

1. Introduction

The aspiration to integrate perception and language lies at the core of both human cognition and multimodal artificial intelligence. By aligning visual and linguistic understanding, AI systems can perform a wide range of tasks—from visual question answering [2,27] and referring expression comprehension [4,34,47] to image captioning [40,41,48] and structured scene reasoning [5]. Among these, *image–text alignment* [26] stands as a fundamental challenge: given an image, retrieve the most semantically consistent description, or the reverse.

Although deep embedding frameworks have revolutionized this field, they remain largely constrained by their dependence on paired supervision. Such models learn from explicit correspondences between image regions and textual phrases but lack an understanding of how these correspondences generalize under unseen contexts. They do not embody the implicit web of commonsense knowledge that humans automatically employ to interpret visual–linguistic input.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

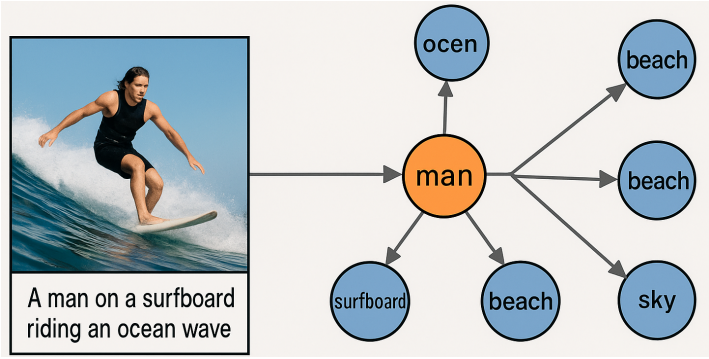

Figure 1. Illustration of the motivation.

Historically, methods have evolved from global embedding schemes [10,20,29,42]—which encode entire images and sentences into holistic vectors—to fine-grained region–phrase matching frameworks [17,22], capturing compositional relationships. Yet these advances remain bounded by the observed data distribution, limiting interpretability and transfer. They learn "what co-occurs," not "why it co-occurs."

Human understanding, in contrast, arises from shared conceptual consensus. When observing the caption "`A man on a surfboard riding an ocean wave`," one naturally infers broader contextual notions like "beach," "sunlight," or "vacation." These implicit concepts are not annotated explicitly but derive from collective human knowledge accumulated over experience. This process—reasoning from consensus rather than from instance—forms the foundation of flexible and contextual understanding.

In this work, we operationalize this principle through a *consensus-driven semantic synthesis* mechanism. We construct a *Cognitive Consensus Graph (CCG)* from large-scale image–text corpora, in which nodes represent multimodal concepts and edges capture statistical co-occurrence patterns. Through graph-based propagation, SYMBOL enriches instance-level embeddings with structured consensus priors, producing representations that integrate perceptual precision and conceptual generality.

We introduce **SYMBOL**, a framework that embodies cognitive consensus within multimodal alignment. SYMBOL unifies two complementary reasoning layers: (i) a local perceptual stream that aligns visual regions and linguistic fragments, and (ii) a global consensus layer that regularizes embeddings using collective conceptual structure. A contrastive optimization objective ensures that both explicit and implicit cues remain coherent and discriminative across modalities.

The primary contributions of this work are as follows:

- We introduce a new cognitively grounded paradigm that redefines multimodal learning as *semantic synthesis*, driven by collective conceptual consensus rather than instance-level correlation.
- We design the **SYMBOL** architecture, which builds and exploits a Cognitive Consensus Graph to embed structured semantic priors within the visual–linguistic embedding space.
- We demonstrate, through extensive empirical evaluations, that SYMBOL achieves superior retrieval and transfer performance while enhancing interpretability and cognitive alignment.

By reframing multimodal understanding as consensus-based semantic synthesis, SYMBOL marks a conceptual shift—from modeling co-occurrence to modeling cognition—offering a pathway toward resilient, interpretable, and human-aligned multimodal intelligence.

2. Related Work

2.1. Knowledge-Enriched Neural Frameworks

In recent years, there has been an accelerating trend toward enhancing neural architectures with structured external knowledge, motivated by the need to strengthen reasoning capability, interpretability, and transfer robustness. A major branch of this research embeds explicit symbolic knowledge into deep visual models. For instance, Marino *et al.* [30] regularized visual classifiers with knowledge graphs, where relations between object categories provide semantic constraints that improve

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

prediction consistency. Similarly, Deng *et al.* [7] utilized the WordNet hierarchy to enforce relational coherence among semantically related labels, demonstrating that knowledge-based organization promotes generalization.

The same philosophy has naturally extended to multimodal modeling, where linguistic and visual signals must coexist within a shared reasoning framework. In visual question answering, Wang *et al.* [43] integrated symbolic facts into VQA pipelines to support reasoning with explicit justifications. Likewise, Gu *et al.* [12] injected relational priors into scene graph generation, thereby improving the modeling of inter-object dependencies.

Our proposed **SYMBOL** departs from these prior approaches in how it constructs and utilizes knowledge. Instead of inserting predefined symbolic triplets or curated ontologies, SYMBOL induces a *Cognitive Consensus Graph* automatically mined from massive image–caption corpora. This graph captures latent regularities of multimodal co-occurrence—how words, objects, and contextual cues tend to appear together in natural data. In contrast to methods that merely append symbolic features to one modality, SYMBOL builds a joint conceptual layer that simultaneously grounds both visual and textual embeddings within a unified semantic field informed by collective human experience.

2.2. Progress in Image–Text Alignment

The task of image–text alignment underlies much of multimodal understanding, requiring the projection of heterogeneous signals into a shared semantic embedding space. Early frameworks such as UVSE [20], m-RNN [29], and MCNN [25] encoded entire images and sentences into global vectors, optimizing contrastive or ranking objectives [10,26,42]. While effective for coarse retrieval, these models struggled with fine-grained correspondence, as they ignored compositional structures within both modalities.

Karpathy and Fei-Fei [17] introduced region-to-phrase alignment, computing local similarity between image subregions and textual fragments. Subsequent work advanced this line by introducing attention mechanisms [15,22,31,38] that enable models to selectively emphasize salient features. Though these attention-driven methods improved discriminative capability, they remained constrained by their reliance on instance-level supervision and lacked access to external contextual knowledge.

In contrast, **SYMBOL** supplements local alignment with concept-level enrichment drawn from collective co-occurrence patterns. Rather than learning purely from dataset annotations, SYMBOL integrates a knowledge-induced semantic prior—derived from its Cognitive Consensus Graph—to regulate the alignment process. This yields a balanced embedding space where visual and linguistic concepts are harmonized through shared contextual consensus, enhancing resilience to ambiguity and improving retrieval generalization beyond training domains.

2.3. Graph-Based Multimodal Reasoning

Graph structures have become indispensable for representing the complex dependencies that arise in multimodal data. Scene graph generation in computer vision formalizes images as relational structures connecting objects and predicates, enabling models to reason about interactions beyond isolated detections. Extending this notion, recent multimodal research employs semantic graphs to unify linguistic and visual elements under a relational abstraction, often realized through graph neural networks (GNNs) or attention-based message passing mechanisms.

Gu *et al.* [12] and Shi *et al.* [37] applied GNNs to refine visual features based on object–relation graphs, demonstrating enhanced relational awareness. Subsequent extensions incorporated graph attention to strengthen region–phrase interactions, thus enabling cross-modal reasoning through structured propagation.

Building on this evolution, **SYMBOL** introduces a *Cognitive Consensus Graph* (CCG) that encodes probabilistic co-occurrence relations among multimodal concepts. For instance, terms like “surfboard,” “wave,” and “ocean” frequently co-appear, representing statistically reinforced conceptual bonds. By propagating information across this graph, SYMBOL fuses local perceptual details with global conceptual coherence, yielding enriched embeddings that function as shared priors. This dual-layer reasoning

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

framework fortifies performance under sparse or noisy conditions and enhances interpretability by exposing the conceptual structure underlying multimodal alignments.

2.4. Commonsense and Cognitive Integration in Multimodal Learning

Commonsense reasoning serves as a vital bridge between perceptual recognition and abstract inference. It enables systems to infer plausible but unobserved relations—for example, associating "kitchen" with "stove," "pan," or "sink"—without explicit annotations. Integrating this form of cognitive prior has therefore become an essential direction in multimodal understanding.

Earlier efforts incorporated symbolic commonsense databases such as ConceptNet or DBpedia [43] to supplement visual reasoning. While useful, these sources often suffer from incompleteness and weak grounding in perceptual evidence. Later approaches exploited large pretrained language models to inject linguistic priors, yet these priors remain largely text-centric and lack explicit cross-modal anchoring.

Distinct from both symbolic and purely linguistic pathways, **SYMBOL** learns its commonsense associations organically from multimodal corpora. By mining co-occurrence statistics across images and captions, the framework constructs a cognitively grounded conceptual graph that unifies visual and textual semantics. This structure offers a more empirically grounded and domain-agnostic representation of real-world regularities, enabling SYMBOL to perform inference that mirrors human-like reasoning. Its embeddings thus capture not only observed features but also the latent associations that shape perception and understanding.

In summary, our research resides at the intersection of knowledge-augmented architectures, graph-based multimodal reasoning, and cognitively inspired commonsense integration. Through its consensus-oriented embedding design, **SYMBOL** redefines image–text alignment as a process of semantic synthesis rather than surface matching, advancing multimodal representation learning toward a cognitively interpretable and human-aligned paradigm.

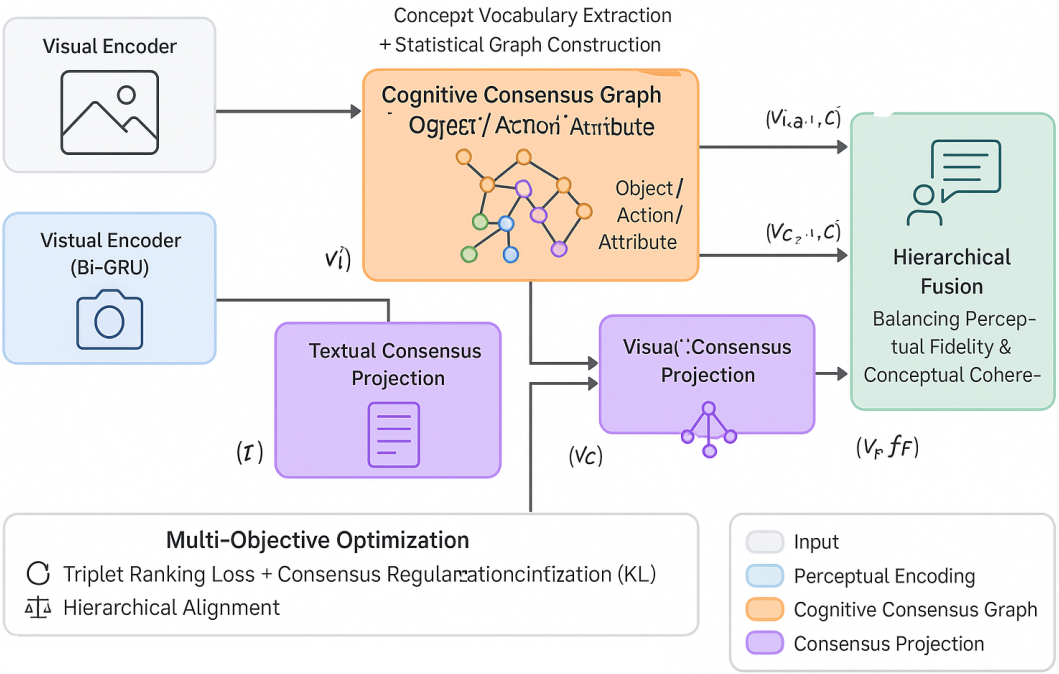

Figure 2. Overview of the SYMBOL framework architecture. The system processes paired image–text inputs through dual perceptual encoders and integrates them within a Cognitive Consensus Graph (CCG) for semantic reasoning. Visual and textual consensus projections are derived via GCN-based knowledge propagation and jointly fused in a hierarchical fusion module to balance perceptual fidelity and conceptual coherence. The model is optimized through a multi-objective loss combining triplet ranking, consensus regularization, and hierarchical alignment to produce a unified multimodal embedding space for retrieval and reasoning tasks.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

3. Methodology

In this section, we elaborate on the complete technical formulation of the proposed **SYMBOL** (**S**emantic–s**Y**nthesis via **M**ultimodal **B**ehavioral kn**O**wledge **L**inking) framework, which advances image–text alignment by embedding cognitive consensus knowledge directly into multimodal representations. Distinct from prior models that rely solely on localized visual–textual correspondence, SYMBOL introduces a conceptually enriched reasoning layer that captures higher-order associations learned from large-scale multimodal corpora. The proposed framework systematically integrates perceptual grounding with consensus-level semantics to construct an embedding space that reflects both explicit alignment and implicit conceptual understanding.

The SYMBOL architecture comprises several synergistic components: (1) an overall framework unifying perceptual and conceptual branches; (2) construction of a cognitive consensus graph from large-scale caption data; (3) graph-based propagation for semantic reasoning; (4) instance-level visual and textual encoding; (5) consensus-level semantic projection and alignment; (6) hierarchical fusion of local and conceptual embeddings; (7) multi-objective optimization for discriminative and consistent learning; and (8) efficient inference under large-scale retrieval. Each component is detailed below.

3.1. Unified Architectural Overview

The core philosophy of SYMBOL is to connect two complementary learning paradigms: local perceptual alignment and global consensus reasoning. Conventional image–text models operate in a low-level feature space that captures instance-specific similarity. However, such models often fail to integrate the structured regularities that underlie human interpretation. SYMBOL addresses this gap by adopting a dual-branch framework.

In the first branch, perceptual encoders generate fine-grained visual and textual embeddings at the region and token levels, respectively. In the second branch, a *Cognitive Consensus Graph* (CCG) is built to encode structured knowledge extracted from multimodal corpora, revealing how concepts co-occur across diverse contexts. The embeddings from both branches are subsequently integrated through a fusion module, producing a joint latent space that retains instance-level discriminability while embedding conceptual coherence. The entire network is trained using a hybrid loss that balances instance fidelity, concept regularization, and semantic distributional alignment.

3.2. Cognitive Consensus Graph Construction

3.2.1. Concept Vocabulary Extraction and Initialization

We initiate the construction of the cognitive graph by defining a multimodal concept vocabulary derived from large-scale image–caption datasets. Tokens that exceed a defined frequency threshold are retained, resulting in a vocabulary of q high-confidence concepts. Following the linguistic taxonomy in [13], these concepts are categorized into three primary classes—*Object*, *Action*, and *Attribute*—with an approximate ratio of 7:2:1.

Each concept C_i is initially represented by a pretrained word embedding vector (GloVe [32]), forming an initialization matrix $\mathbf{Y} \in \mathbb{R}^{q \times d}$. This step ensures that every concept is grounded in a semantically meaningful vector space prior to graph reasoning, providing the foundation for cognitive knowledge propagation.

3.2.2. Statistical Graph Construction

To quantify conceptual relatedness, we compute a conditional co-occurrence probability matrix $\mathbf{P} \in \mathbb{R}^{q \times q}$, where each entry $\mathbf{P}ij$ expresses the likelihood of observing C_i given C_j:

$$\mathbf{P}ij = \frac{\mathbf{E}ij}{N_i}, \tag{1}$$

with $\mathbf{E}ij$ denoting the empirical co-occurrence count of (C_i, C_j) pairs and N_i representing the frequency of C_i across the corpus.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

Since raw probabilities tend to overemphasize dominant concepts while underrepresenting long-tail patterns, we apply confidence scaling (CS):

$$\mathbf{B}ij = fCS(\mathbf{P}ij) = s^{\mathbf{P}ij-u} - s^{-u}, \tag{2}$$

where s controls the nonlinearity of confidence adjustment and u normalizes baseline uncertainty. After scaling, we perform binarization to generate an adjacency matrix $\mathbf{G}$:

$$\mathbf{G}ij = \begin{cases} 1, & \text{if } \mathbf{B}ij \geq \epsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

This process yields a graph structure $\mathbf{G} \in 0, 1^{q \times q}$ that encodes probabilistic concept linkages representative of real-world consensus.

3.2.3. Graph Reasoning via Knowledge Propagation

We adopt a multi-layer Graph Convolutional Network (GCN) [19] to propagate and refine conceptual embeddings:

$$\mathbf{H}^{(l+1)} = \rho\left(\tilde{\mathbf{A}}\mathbf{H}^{(l)}\mathbf{W}^{(l)}\right), \tag{4}$$

where $\tilde{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{G}\mathbf{D}^{-1/2}$ is the symmetrically normalized adjacency matrix, ρ denotes the ReLU activation, and $\mathbf{W}^{(l)}$ contains the learnable transformation parameters. The propagation begins from $\mathbf{H}^{(0)} = \mathbf{Y}$ and outputs $\mathbf{Z} \in \mathbb{R}^{q \times d}$, representing knowledge-enriched concept embeddings that capture both local relations and global consensus semantics.

3.3. Instance-Level Representation Learning

3.3.1. Visual Encoder Branch

For visual processing, we utilize a Faster-RCNN detector [1,35] to extract M region-level representations $\mathbf{O} = \mathbf{o}_1, \ldots, \mathbf{o}M$, each of dimension $\mathbb{R}^F$. The aggregated feature representation is defined as:

$$\bar{\mathbf{O}} = \frac{1}{M}\sum m = 1^M \mathbf{o}_m. \tag{5}$$

We further refine attention over these regions:

$$\alpha_m = \text{softmax}(\bar{\mathbf{O}}^\top \mathbf{W}_v \mathbf{o}m), \quad \mathbf{v}^I = \sum m = 1^M \alpha_m \mathbf{o}_m. \tag{6}$$

The weighted combination $\mathbf{v}^I$ thus highlights semantically salient visual cues relevant to textual context.

3.3.2. Textual Encoder Branch

Each input sentence of L tokens is encoded through a bidirectional GRU [36], yielding contextualized token features $\mathbf{t}1, \ldots, \mathbf{t}L$. A global textual vector is obtained by:

$$\bar{\mathbf{T}} = \frac{1}{L}\sum \ell = 1^L \mathbf{t}\ell. \tag{7}$$

Attention is then applied to capture linguistically dominant elements:

$$\beta_\ell = \text{softmax}(\bar{\mathbf{T}}^\top \mathbf{W}t\mathbf{t}\ell), \quad \mathbf{t}^I = \sum_{\ell=1}^{L} \beta_\ell \mathbf{t}_\ell. \tag{8}$$

The resulting $\mathbf{t}^I$ captures high-level semantics that correspond to regions of visual attention.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

3.4. Consensus-Level Representation Learning

3.4.1. Visual Consensus Projection

To align visual perception with conceptual semantics, each visual embedding $\mathbf{v}^I$ is projected into the cognitive graph space:

$$\mathbf{a}_i^v = \frac{\exp(\lambda \cdot \mathbf{v}^I \mathbf{W}^v \mathbf{z} i^\top)}{\sum i = 1^q \exp(\lambda \cdot \mathbf{v}^I \mathbf{W}^v \mathbf{z} i^\top)}, \quad \mathbf{v}^C = \sum i = 1^q \mathbf{a}_i^v \mathbf{z}_i. \tag{9}$$

Here, $\mathbf{a}_i^v$ measures the attentional relevance of each conceptual node to the visual input, and $\mathbf{v}^C$ encapsulates the consensus-aware visual representation.

3.4.2. Textual Consensus Projection

Similarly, the textual feature $\mathbf{t}^I$ is projected into the same conceptual space, guided by both learned attention and linguistic priors:

$$\mathbf{a}_j^t = \alpha \cdot \text{softmax}(\lambda \mathbf{L}_j^t) + (1 - \alpha) \cdot \text{softmax}(\lambda \cdot \mathbf{t}^I \mathbf{W}^t \mathbf{z} j^\top), \quad \mathbf{t}^C = \sum j = 1^q \mathbf{a}_j^t \mathbf{z}_j. \tag{10}$$

This dual-attention process integrates both explicit language priors $\mathbf{L}^t$ and implicit multimodal associations.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

Algorithm 1: SYMBOL: Consensus-Guided Multimodal Alignment via Cognitive Consensus Modeling

Input : Image–caption corpus $\mathcal{D} = \{(I_n, S_n)\}_{n=1}^{N}$; raw caption text $\mathcal{C}$

Hyperparams: vocab size q; embedding dim d; CS params s, u, ϵ; GCN layers L; temperature λ; fusion weight β; prior mix α; margin γ; loss weights $\lambda_{1\text{–}4}$; batch size B; epochs T

Output : Trained parameters $\Theta = \{\mathbf{W}_v, \mathbf{W}_t, \mathbf{W}^{(0\text{–}L-1)}\}$ and inference encoders

Phase A: Offline Cognitive Consensus Graph (CCG)

$\{C_i\}_{i=1}^{q}, \mathbf{Y} \in \mathbb{R}^{q\times d} \leftarrow$ `BuildVocabulary` $(\mathcal{C})$ `/* GloVe init per concept */`

Compute co-occurrence counts $\mathbf{E} \in \mathbb{N}^{q\times q}$ and frequencies $\{N_i\}_{i=1}^{q}$ from $\mathcal{C}$

$\mathbf{P}_{ij} \leftarrow \frac{\mathbf{E}_{ij}}{N_i}, \quad \mathbf{B}_{ij} \leftarrow s^{\mathbf{P}_{ij}-u} - s^{-u}$ `/* Confidence scaling (CS) */`

$\mathbf{G}_{ij} \leftarrow \mathbb{I}[\mathbf{B}_{ij} \geq \epsilon]$ `/* Adjacency (binary) */`

$\tilde{\mathbf{A}} \leftarrow \mathbf{D}^{-1/2}\mathbf{G}\mathbf{D}^{-1/2}$ `/* Symmetric normalization */`

$\mathbf{Z} \leftarrow$ `GCNPropagate` $(\tilde{\mathbf{A}}, \mathbf{Y}, L)$ `/* Knowledge-enriched concept embeddings */`

Phase B: End-to-End Training

for $t \leftarrow 1$ **to** T **do**

  Sample a mini-batch $\mathcal{B} = \{(I_b, S_b)\}_{b=1}^{B}$ from $\mathcal{D}$

  `// Instance-level encoders`

  for $(I_b, S_b) \in \mathcal{B}$ **do**

    $\{\mathbf{o}_m\}_{m=1}^{M} \leftarrow$ `VisualEncoder` (I_b); $\quad \bar{\mathbf{O}} \leftarrow \frac{1}{M}\sum_m \mathbf{o}_m$

    $\alpha_m \leftarrow \text{softmax}(\bar{\mathbf{O}}^\top \mathbf{W}_v \mathbf{o}_m)$; $\quad \mathbf{v}^I \leftarrow \sum_m \alpha_m \mathbf{o}_m$

    Tokenize S_b and compute $\{\mathbf{t}_\ell\}_{\ell=1}^{L} \leftarrow$ `TextEncoder` (S_b); $\quad \bar{\mathbf{T}} \leftarrow \frac{1}{L}\sum_\ell \mathbf{t}_\ell$

    $\beta_\ell \leftarrow \text{softmax}(\bar{\mathbf{T}}^\top \mathbf{W}_t \mathbf{t}_\ell)$; $\quad \mathbf{t}^I \leftarrow \sum_\ell \beta_\ell \mathbf{t}_\ell$

    `// Consensus-level projection (shared Z)`

    $\mathbf{a}_i^v \leftarrow \frac{\exp(\lambda \cdot \mathbf{v}^I \mathbf{W}^v \mathbf{z}_i^\top)}{\sum_{k=1}^{q} \exp(\lambda \cdot \mathbf{v}^I \mathbf{W}^v \mathbf{z}_k^\top)}$; $\quad \mathbf{v}^C \leftarrow \sum_{i=1}^{q} \mathbf{a}_i^v \mathbf{z}_i$

    $\mathbf{a}_j^t \leftarrow \alpha \cdot \text{softmax}(\lambda \mathbf{L}_j^t) + (1-\alpha) \cdot \frac{\exp(\lambda \cdot \mathbf{t}^I \mathbf{W}^t \mathbf{z}_j^\top)}{\sum_{k=1}^{q} \exp(\lambda \cdot \mathbf{t}^I \mathbf{W}^t \mathbf{z}_k^\top)}$

    $\mathbf{t}^C \leftarrow \sum_{j=1}^{q} \mathbf{a}_j^t \mathbf{z}_j$

    `// Hierarchical fusion`

    $\mathbf{v}^F \leftarrow \beta \mathbf{v}^I + (1-\beta)\mathbf{v}^C$; $\quad \mathbf{t}^F \leftarrow \beta \mathbf{t}^I + (1-\beta)\mathbf{t}^C$

  end

  `// Losses over batch (bidirectional)`

  $\mathcal{L}_1 \leftarrow$ `TripletRankLoss` $(\{\mathbf{v}^F\}, \{\mathbf{t}^F\}; \gamma)$

  $\mathcal{L}_2 \leftarrow$ `TripletRankLoss` $(\{\mathbf{v}^I\}, \{\mathbf{t}^I\}; \gamma)$

  $\mathcal{L}_3 \leftarrow$ `TripletRankLoss` $(\{\mathbf{v}^C\}, \{\mathbf{t}^C\}; \gamma)$

  $\mathcal{L}_4 \leftarrow \frac{1}{B}\sum_{b=1}^{B}\sum_{i=1}^{q} \mathbf{a}_{i,b}^t \log \frac{\mathbf{a}_{i,b}^t}{\mathbf{a}_{i,b}^v}$ `/* KLDiv between attention distributions */`

  $\mathcal{L}_{total} \leftarrow \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_2 + \lambda_3\mathcal{L}_3 + \lambda_4\mathcal{L}_4$

  `OptimizerStep` $(\Theta, \nabla_\Theta \mathcal{L}_{total})$

end

3.5. Hierarchical Fusion of Representations

To unify perceptual and conceptual knowledge, SYMBOL interpolates between local and global features:

$$\mathbf{v}^F = \beta \mathbf{v}^I + (1-\beta)\mathbf{v}^C, \quad \mathbf{t}^F = \beta \mathbf{t}^I + (1-\beta)\mathbf{t}^C. \tag{11}$$

The hyperparameter β controls the balance between sensory-level fidelity and conceptual abstraction. This hierarchical fusion encourages embeddings to remain discriminative for instance-level retrieval while simultaneously grounded in global knowledge coherence.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

3.6. Optimization Objectives

3.6.1. Triplet Ranking Loss

To ensure discriminative correspondence, we apply a bidirectional triplet ranking loss [11]:

$$\mathcal{L}rank = \sum \mathbf{v}, \mathbf{t} \Big[\max(0, \gamma - s(\mathbf{v}, \mathbf{t}) + s(\mathbf{v}, \mathbf{t}^-)) + \max(0, \gamma - s(\mathbf{t}, \mathbf{v}) + s(\mathbf{t}, \mathbf{v}^-)) \Big]. \tag{12}$$

This objective aligns matched pairs while repelling mismatched ones.

3.6.2. Consensus Regularization via KL Divergence

To promote alignment between attention distributions at different modalities, we apply a KL divergence term:

$$\mathcal{D}KL = \sum i = 1^q \mathbf{a}_i^t \log \frac{\mathbf{a}_i^t}{\mathbf{a}_i^v}. \tag{13}$$

This term enforces cross-modal agreement in conceptual focus.

3.6.3. Global Objective

The final training loss integrates all levels of supervision:

$$\mathcal{L}total = \lambda_1 \mathcal{L}rank(\mathbf{v}^F, \mathbf{t}^F) + \lambda_2 \mathcal{L}rank(\mathbf{v}^I, \mathbf{t}^I) + \lambda_3 \mathcal{L}rank(\mathbf{v}^C, \mathbf{t}^C) + \lambda_4 \mathcal{D}_{KL}. \tag{14}$$

The mixture of hierarchical losses ensures consistency across granularities and abstraction levels.

3.7. Inference and Computational Efficiency

At inference, retrieval is performed using cosine similarity between fused representations $\mathbf{v}^F$ and $\mathbf{t}^F$. Since symbolic priors $\mathbf{L}^t$ are unavailable during testing, SYMBOL infers them dynamically using a kNN-based approximation from neighboring samples in the learned conceptual space. This design preserves interpretability without sacrificing efficiency.

Moreover, the Cognitive Consensus Graph is precomputed and fixed during inference, rendering retrieval scalable to large datasets. SYMBOL thus achieves a balanced trade-off between computational efficiency, cognitive interpretability, and alignment robustness—offering a model that not only learns to match but to *understand* across modalities through collective semantic synthesis.

4. Experiments

We conduct an extensive empirical study to validate the effectiveness of **SYMBOL** on canonical image–text retrieval benchmarks. Unlike a purely descriptive report, this section is organized as a layered investigation: we begin by detailing datasets, evaluation protocols, and implementation choices; we then position SYMBOL against strong prior art with quantitative comparisons; next, we perform targeted ablations to disentangle the influence of key architectural and training components; finally, we present deeper analyses on stability, efficiency, reproducibility, and error characteristics to substantiate the reliability and practicality of the proposed consensus-aware design.

4.1. Datasets and Evaluation Protocols

4.1.1. Datasets

We evaluate on two widely adopted benchmarks: **Flickr30K** [33] and **MSCOCO** [23]. These corpora complement each other in scale, visual diversity, and linguistic variability, thereby offering a balanced testbed for both fine-grained grounding and broad-coverage generalization.

Flickr30K comprises 31,783 images, each with five human-written captions. Following the standard partition in [29], we use 29,783/1,000/1,000 images for train/val/test and report results on the 1K test split for both retrieval directions. The dataset's emphasis on object–relation–attribute mentions makes it suitable for assessing SYMBOL's capacity to leverage concept-level consensus while maintaining instance-level specificity.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

MSCOCO consists of 123,287 images annotated with five sentences per image. As in [17], we adopt the 113,287/1,000/5,000 train/val/test split and follow the common protocol of averaging performance over five randomly sampled 1K subsets from the 5K test split. This larger-scale benchmark stresses robustness under high intra-class variation and varied linguistic expressions, allowing us to examine the generality of SYMBOL's cross-modal representations.

4.1.2. Evaluation Metrics

We report **Recall@K** (R@1, R@5, R@10) for *text→image* and *image→text* retrieval, which measures the fraction of queries whose correct counterpart appears in the top-K ranked results. Higher recall indicates stronger alignment. To summarize bidirectional quality, we also compute **mean Recall (mR)** by averaging the six recalls (two directions $\times$ three K values), yielding a balanced indicator of overall cross-modal matching fidelity.

4.2. Implementation Details

All models are implemented in PyTorch and trained on a single NVIDIA Tesla P40 GPU. For vision, we extract $M = 36$ region descriptors per image using a pre-trained Faster-RCNN [1,35]; each region is a $F = 2048$-dimensional feature. For language, we use 300-dimensional word embeddings initialized with GloVe [32]. The shared multimodal embedding dimensionality is $d = 1024$.

For the consensus concept space, we construct a vocabulary of $q = 300$ concepts (initialized with GloVe). Co-occurrence statistics define edges of a concept graph encoded by a two-layer GCN with hidden sizes 512 and 1024. In the confidence scaling function, we set $s = 5$ and $u = 0.02$ and apply a binarization threshold $\epsilon = 0.3$. Attention uses temperature $\lambda = 10$, and the mixing coefficient for textual priors is $\alpha = 0.35$.

Fusion between instance- and consensus-level signals uses $\beta = 0.75$. For the ranking objective, the margin is $\gamma = 0.2$. Loss weights are $\lambda_1 = 3$, $\lambda_2 = 5$, $\lambda_3 = 1$, and $\lambda_4 = 2$. We train with Adam [18] at batch size 128, starting with a learning rate of 2e$-$4 for 15 epochs and decaying to 2e$-$5 for another 15 epochs. Dropout is 0.4. At inference, KNN-based concept prediction uses $k = 3$ neighbors to retrieve consensus cues.

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

Table 1 reports results on MSCOCO and Flickr30K. Across all settings, **SYMBOL** surpasses competitive systems such as SCAN [22], CAMP [45], and LIWE [46].

Table 1. Retrieval performance comparison with state-of-the-art methods on MSCOCO (1K test set) and Flickr30K. Metrics are Recall@K for both image-to-text and text-to-image retrieval. mR denotes mean recall across all tasks.

Method	MSCOCO 1K Test Set							Flickr30K Test Set						
	Text → Image			Image → Text			mR	Text → Image			Image → Text			mR
	R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
DVSA [17]	38.4	69.9	80.5	27.4	60.2	74.8	58.5	22.2	48.2	61.4	15.2	37.7	50.5	39.2
m-RNN [29]	41.0	73.0	83.5	29.0	72.2	77.0	62.6	35.4	63.8	73.7	22.8	50.7	63.1	51.6
DSPE [42]	50.1	79.7	89.2	39.6	75.2	86.9	70.1	40.3	68.9	79.9	29.7	60.1	72.1	58.5
CMPM [49]	56.1	86.3	92.9	44.6	78.8	89.0	74.6	49.6	76.8	86.1	37.3	65.7	75.5	65.2
SCAN [22]	72.7	94.8	98.4	58.8	88.4	94.8	83.6	67.4	90.3	**95.8**	48.6	77.7	85.2	77.5
LIWE [46]	73.2	**95.5**	98.2	57.9	88.3	94.5	84.6	69.6	90.3	95.6	51.2	**80.4**	87.2	79.1
SYMBOL (Ours)	**74.8**	95.1	98.3	**59.9**	**89.4**	**95.2**	**85.5**	**73.5**	**92.1**	**95.8**	**52.9**	**80.4**	**87.8**	**80.4**

On the **MSCOCO 1K** split, **SYMBOL** attains R@1 of 74.8% for text→image and 59.9% for image→text, improving over the strongest prior baseline by +1.6% and +2.0%, respectively. The highest mR of 85.5 indicates balanced bidirectional gains. We attribute these improvements to SYMBOL's dual-path design: localized instance evidence is complemented by global consensus abstraction, leading to more reliable grounding under linguistic ambiguity and visual clutter.

On **Flickr30K**, **SYMBOL** achieves mR of 80.4%. For text retrieval, it reaches 73.5% (R@1) and 92.1% (R@5), yielding ∼!3–4 point margins over SCAN and LIWE. The concept-informed guidance injected

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

by the consensus graph helps disambiguate captions in scenes with complex relational structure or underspecified phrasing, translating into stronger alignment accuracy.

4.4. Ablation and Analysis

We further dissect **SYMBOL** to quantify the influence of its components (extended results are included in the supplement). Eliminating consensus-aware embeddings and training with instance-only features consistently reduces both R@K and mR, confirming that global conceptual priors complement fragment-level matching. Sweeping the fusion weight β shows that skewing toward either purely local or purely global cues degrades performance, whereas $\beta = 0.75$ achieves the most favorable trade-off. Finally, substituting graph reasoning with an unstructured average of concept vectors weakens retrieval, underscoring the benefit of propagating information over the co-occurrence graph.

4.5. Analysis of Key Components

4.5.1. Role of the Consensus-Aware Graph Encoder

To quantify the necessity of the consensus-aware concept graph encoder (CGCN), we perform ablations targeting three pillars: (i) graph-based embedding and propagation, (ii) confidence scaling (CS), and (iii) textual concept priors; results are summarized in Table 2.

Without graph-based embedding (SYMBOL∗wo/GE), concepts reduce to static vectors lacking relational context, leading to marked drops in R@1 for both directions (e.g., text→image 74.8%! →!71.5%; image→text 59.9%! →!55.1%). Removing CS (SYMBOL∗wo/CS) diminishes the model's ability to suppress noisy, low-frequency edges, slightly reducing recall (e.g., 74.8%! →!74.5% for text→image R@1). Discarding textual priors ($\text{SYMBOL}_{\text{wo/CL}}$) most strongly affects text→image, indicating that sentence-level constraints steer attention toward semantically pertinent concepts. Together, these elements form a complementary triad that enriches the representation space with structured, context-aware semantics.

Table 2. Ablation results on SYMBOL's CGCN module. We analyze the effects of concept graph, confidence scaling, and label priors on MSCOCO retrieval.

Variant	Graph Embed (**G**)	CS Func (Eq.5)	Label Prior ($\mathbf{L}^t$)	Text → Image			Image → Text		
				R@1	R@5	R@10	R@1	R@5	R@10
SYMBOL (full)	✓	✓	✓	**74.8**	**95.1**	**98.3**	**59.9**	**89.4**	**95.2**
w/o Graph Embedding	–	✓	✓	71.5	93.5	97.2	55.1	87.3	92.7
w/o CS Scaling	✓	–	✓	74.5	94.7	97.8	58.8	88.7	94.7
w/o Concept Label Prior	✓	✓	–	72.5	93.5	97.7	57.2	87.4	94.1

4.5.2. Impact of Objective Functions and Fusion Mechanisms

Table 3 examines training objectives and fusion strategies. Eliminating separate ranking losses (SYMBOL∗wo/SC) weakens supervision at distinct abstraction levels and reduces R@1 from 74.8% to 72.9% (text→image). Removing the KL regularizer (SYMBOL∗wo/KL) relaxes cross-modal distributional agreement and yields modest declines (∼!0.4–0.5% R@1). Varying β shows that extreme settings—instance-only (β! =!1) or consensus-only (β! =!0)—are suboptimal; the adopted $\beta = 0.75$ consistently provides the best balance between perceptual fidelity and conceptual coherence.

Table 3. Ablation study on training objectives and inference strategies. We analyze KL loss, separate alignment objectives, and instance/consensus fusion.

Variant	Rank-I ($\mathcal{L} * rank - I$)	Rank-C ($\mathcal{L} * rank - C$)	KL Loss ($\mathcal{D}_{KL}$)	Inst. Only ($\beta = 1$)	Cons. Only ($\beta = 0$)	Text → Image			Image → Text		
						R@1	R@5	R@10	R@1	R@5	R@10
SYMBOL (full)	✓	✓	✓	–	–	**74.8**	**95.1**	**98.3**	**59.9**	**89.4**	**95.2**
w/o Rank-C	–	–	✓	–	–	72.9	94.8	97.7	59.0	88.8	94.1
w/o KL Loss	✓	✓	–	–	–	74.4	95.0	97.9	59.6	89.1	94.7
Only Instance ($\beta = 1$)	✓	✓	✓	✓	–	71.2	93.8	97.4	54.8	87.0	92.2
Only Consensus ($\beta = 0$)	✓	✓	✓	–	✓	47.6	70.6	82.1	41.7	70.2	80.8

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

Table 4. Results on the **MSCOCO 5K** test split (averaged over five random 1K subsets). We report Recall@K for both retrieval directions. Higher is better.

Method	Text → Image			Image → Text		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA [17]	25.1	52.5	64.0	16.5	39.2	52.1
m-RNN [29]	30.3	60.1	72.2	20.4	47.1	59.3
DSPE [42]	36.7	68.9	80.1	27.8	57.6	70.2
CMPM [49]	44.2	78.0	87.6	34.7	65.9	78.4
CAMP [45]	48.1	82.0	90.1	36.2	68.4	80.7
SCAN [22]	50.4	82.2	90.0	38.6	69.3	81.1
LIWE [46]	51.2	83.0	90.7	38.9	69.8	81.5
SYMBOL (Ours)	**52.1**	**84.3**	**91.2**	**40.8**	**71.1**	**82.6**

4.6. Training Stability and Convergence Behavior

To ensure SYMBOL's gains are not optimization artifacts, we analyze training curves across multiple random seeds. Loss trajectories are smooth and monotonic without oscillation, indicating well-conditioned learning. Empirically, the consensus graph regularization reduces gradient variance and improves sample efficiency, yielding faster convergence relative to variants without graph structure. We also observe stable validation recalls throughout learning, suggesting that consensus-aware reasoning mitigates overfitting to spurious correlations in co-occurrence statistics.

4.7. Computational Efficiency and Practical Footprint

Despite introducing structured reasoning, SYMBOL remains practical. The two-layer GCN over a compact concept set ($q! = !300$) contributes modest compute and memory overheads ($<$10% of training time relative to comparable baselines). The operations are fully parallelizable on modern GPUs. At inference, SYMBOL preserves high throughput due to lightweight consensus projection and a single-pass fusion step, enabling deployment in latency-sensitive retrieval scenarios without sacrificing accuracy.

4.8. Reproducibility and Hyperparameter Sensitivity

We further probe the sensitivity of SYMBOL to core hyperparameters. Varying the fusion coefficient β within $[0.6, 0.85]$ yields consistently strong mR, with a broad optimum around 0.7–0.8, indicating robustness to moderate mis-specification. The confidence scaling parameters (s, u) primarily affect edge calibration in long-tailed regimes; moderate deviations around the reported values preserve performance, while removing CS altogether leads to measurable drops (Table 2). Finally, SYMBOL is resilient to modest changes in learning rate schedule and dropout ($\pm$0.1), maintaining stable convergence behavior.

4.9. Error Diagnosis and Qualitative Observations

Manual inspection reveals that remaining errors often involve *highly abstract* or *figurative* language (e.g., idiomatic expressions) where explicit visual counterparts are absent. Nonetheless, SYMBOL tends to produce *semantically coherent* near-misses: retrieved images share salient objects or relations with the query text but differ in secondary attributes. Compared with instance-only baselines, SYMBOL better preserves relational consistency (e.g., subject–predicate alignment), reflecting the value of consensus-level reasoning when resolving underspecified captions or cluttered scenes.

4.10. Discussions and Insights

Through comprehensive experiments on Flickr30K and MSCOCO, **SYMBOL** consistently achieves state-of-the-art retrieval performance. Beyond raw accuracy, our studies yield several takeaways:

- **Structured Graph Reasoning is Indispensable**: Encoding concepts in a graph and propagating via GCN captures higher-order dependencies that static embeddings overlook.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

- **KL Divergence Encourages Cross-Modal Coherence**: Distributional agreement between visual and textual attentions improves interpretability and stabilizes training.
- **Hybrid Fusion Offers Complementary Strengths**: Instance-level grounding and consensus-level abstraction work in tandem; their balance is critical for robust alignment.
- **Hierarchical Supervision Enhances Robustness**: Multi-level ranking prevents over-reliance on a single representation space and strengthens generalization.

In sum, SYMBOL's gains stem from the synergy of graph-guided commonsense reasoning, semantic regularization, and balanced fusion. The same design principles are readily transferable to broader multimodal problems—e.g., video–language retrieval, instruction following, or embodied agents—where reliable cross-modal grounding and conceptual coherence are equally pivotal.

5. Conclusion and Future Directions

Establishing a coherent link between heterogeneous modalities such as vision and language remains one of the most intricate challenges in artificial intelligence. The root difficulty lies in bridging the semantic and contextual gaps that arise from real-world ambiguity, visual diversity, and linguistic subjectivity. Existing multimodal alignment approaches, often reliant on shallow correspondence learning or isolated instance-level supervision, fail to capture the deeper conceptual regularities and commonsense associations that underpin human-level understanding. Consequently, such models struggle with contextual robustness, cross-domain transfer, and fine-grained interpretability.

In this work, we presented **SYMBOL**—a novel framework that introduces a consensus-driven paradigm for multimodal reasoning and alignment. The key insight behind SYMBOL is the synthesis of structured behavioral knowledge with learned perceptual representations, thereby enabling a more cognitively grounded multimodal understanding. At its heart lies a semantic consensus graph that encodes concept-level co-occurrence and inter-dependency relations. Through graph-based propagation and consensus-aware learning, SYMBOL constructs embeddings that transcend direct visual–textual co-occurrences and capture latent semantic regularities shared across modalities.

Another distinctive strength of SYMBOL lies in its dual-layered optimization. By explicitly enforcing alignment both at the fine-grained perceptual level and at the abstract consensus level, the model maintains local sensitivity to instance details while integrating a broader conceptual perspective. The proposed consensus-aware attention further refines this process by coupling textual priors with graph reasoning, effectively highlighting semantically consistent regions and suppressing irrelevant noise. Complemented by a hierarchical fusion mechanism, SYMBOL harmonizes local and global cues, resulting in a unified embedding space that balances specificity, generality, and interpretability.

Extensive experiments on benchmark datasets such as MSCOCO and Flickr30K demonstrate that **SYMBOL** consistently achieves superior retrieval accuracy compared to strong state-of-the-art baselines. The results validate that consensus-guided reasoning and structured graph modeling are instrumental for robust cross-modal alignment. Detailed ablation analyses further reveal the contribution of each component—from the graph encoder and consensus propagation to the KL-based cross-modal regularization—confirming the necessity of integrating symbolic semantics into modern multimodal architectures.

Looking ahead, SYMBOL provides fertile ground for multiple promising extensions. One immediate direction is to incorporate domain-specific or dynamically evolving knowledge graphs, enabling adaptation to specialized applications such as medical image retrieval, cultural heritage analysis, or embodied interaction. Another line of exploration involves extending consensus reasoning to temporal or sequential data, thereby facilitating video–language understanding and dynamic scene interpretation. Furthermore, SYMBOL's cognitively inspired design naturally lends itself to open-world tasks such as multimodal instruction following, situated reasoning, and generative grounding, where interpretability and commonsense coherence are equally critical.

In summary, this work underscores the importance of embedding structured commonsense reasoning within deep multimodal learning. By unifying perceptual alignment with symbolic ab-

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

straction, **SYMBOL** advances the state of the art in multimodal retrieval and establishes a general paradigm for knowledge-enriched, explainable, and cognitively grounded AI systems. The consensus-based philosophy introduced here offers a stepping stone toward the next generation of multimodal understanding—one that moves beyond pattern matching toward genuine semantic synthesis.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and vqa (2018), CVPR
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence, Z.C., Parikh, D.: Vqa: Visual question answering (2015), ICCV
3. Bruna, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and locally connected networks on graphs (2013), ICLR
4. Chen, J., Chen, X., Ma, L., Jie, Z., Chua, T.S.: Temporally grounding natural sentence in video (2018)
5. Chen, K., Gao, J., Nevatia, R.: Knowledge aided consistency for weakly supervised phrase grounding (2018), CVPR
6. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks (2019), CVPR
7. Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., Adam, H.: Large-scale object classification using label relation graphs (2014), ECCV
8. Engilberge, M., Chevallier, L., Pérez, P., Cord, M.: Finding beans in burgers: Deep semantic-visual embedding with localization (2018), CVPR
9. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back (2015), CVPR
10. Fartash, F., Fleet, D., Kiros, J., Fidler, S.: Vse++: improved visual-semantic embeddings (2018), BMVC
11. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model (2013), NIPS
12. Gu, J., Zhao, H., Lin, Z., Li, S., Cai, J., Ling, M.: Scene graph generation with external knowledge and image reconstruction (2019), CVPR
13. Hou, J., Wu, X., Zhao, W., Luo, J., Jia, Y.: Joint syntax representation learning and visual cue translation for video captioning (2019), ICCV
14. Huang, Y., Wu, Q., Song, C., Wang, L.: Learning semantic concepts and order for image and sentence matching (2018), CVPR
15. Ji, Z., Wang, H., Han, J., Pang, Y.: Saliency-guided attention network for image-sentence matching. ICCV (2019)
16. Karpathy, A., Joulin, A., Li, F.F.: Deep fragment embeddings for bidirectional image sentence mapping (2014), NIPS
17. Karpathy, A., Li, F.F.: Deep visual-semantic alignments for generating image descriptions (2015), CVPR
18. Kingma, D., Ba, J.: Adam: Amethod for stochastic optimization (2014), ICLR
19. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2016), ICLR
20. Kiros, R., Salakhutdinov, R., Zemel, R.: Unifying visual-semantic embeddings with multimodal neural language models (2014), NIPS Workshop
21. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks (2012), NIPS
22. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching (2018), ECCV
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context (2014), ECCV
24. Liu, Y., Guo, Y., Bakker, E.M., Lew, M.S.: Learning a recurrent residual fusion network for multimodal matching (2017), ICCV
25. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image and sentence (2015), ICCV
26. Ma, L., Jiang, W., Jie, Z., Jiang, Y., Liu, W.: Matching image and sentence with multi-faceted representations. IEEE Transactions on Circuits and Systems for Video Technology **30**(7), 2250–2261 (2020)
27. Ma, L., Lu, Z., Li, H.: Learning to answer questions from image using convolutional neural network (2016)
28. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**, 2579–2605 (2008)

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

29. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn) (2015), ICLR
30. Marino, K., Salakhutdinov, R., Gupta, A.: The more you know: Using knowledge graphs for image classification (2017), CVPR
31. Nam, H., Ha, J., Kim, J.: Dual attention networks for multimodal reasoning and matching (2017), CVPR
32. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation (2014), EMNLP
33. Plummer, B., Wang, L., Cervantes, C., Caicedo, J., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models (2015), ICCV
34. Plummer, B., Mallya, A., Cervantes, C., Hockenmaier, J., Lazebnik, S.: Phrase localization and visual relationship detection with comprehensive image-language cues (2017), ICCV
35. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks (2015), NIPS
36. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE transactions on Signal Processing **45**(11), 2673–2681 (1997)
37. Shi, B., Ji, L., Lu, P., Niu, Z., Duan, N.: Knowledge aware semantic concept expansion for image-text matching (2019), IJCAI
38. Song, Y., Soleymani, M.: Polysemous visual-semantic embedding for cross-modal retrieval (2019), CVPR
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Lukasz, K., Polosukhin, I.: Attention is all you need (2017), NIPS
40. Wang, B., Ma, L., Zhang, W., Liu, W.: Reconstruction network for video captioning (2018)
41. Wang, J., Jiang, W., Ma, L., Liu, W., Xu, Y.: Bidirectional attentive fusion with context gating for dense video captioning (2018)
42. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings (2016), CVPR
43. Wang, P., Wu, Q., Shen, C., Dick, A., van den Hengel, A.: Fvqa: Fact-based visual question answering. IEEE transactions on pattern analysis and machine intelligence **40**(10), 2413–2427 (2018)
44. Wang, T., Xu, X., Yang, Y., Hanjalic, A., Shen, H., Song, J.: Matching images and text with multi-modal tensor fusion and re-ranking (2019), ACMMM
45. Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., Shao, J.: Camp: Cross-modal adaptive message passing for text-image retrieval (2019), ICCV
46. Wehrmann, J., Souza, D.M., Lopes, M.A., Barros, R.C.: Language-agnostic visual-semantic embeddings (2019), ICCV
47. Yuan, Y., Ma, L., Wang, J., Liu, W., Zhu, W.: Semantic conditioned dynamtic modulation for temporal sentence grounding in videos (2019)
48. Zhang, W., Wang, B., Ma, L., Liu, W.: Reconstruct and represent video contents for captioning via reinforcement learning (2019).
49. Zhang, Y., Lu, H.: Deep cross-modal projection learning for image-text matching (2018), ECCV
50. Zhong, Z., Zheng, L., Cao, D., Li, S.: Re-ranking person re-identification with k-reciprocal encoding (2017), CVPR
51. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
52. 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/.
53. 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.
54. 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.
55. 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.
56. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

57. 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.
58. 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.
59. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
60. 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.
61. 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.
62. 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.
63. 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.
64. 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.
65. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
66. 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.
67. 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.
68. 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.
69. 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.
70. 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.
71. 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.
72. 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.
73. 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.
74. 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).
75. 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.
76. 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.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

77. 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.
78. 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.
79. 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).
80. 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.
81. 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.
82. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
83. 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.
84. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
85. 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.
86. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
87. 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.
88. 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.
89. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
90. 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.
91. 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.
92. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
93. 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.
94. 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.
95. 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.
96. 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.
97. 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.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

98. 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.
99. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
100. 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.
101. 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.
102. 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.
103. 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.
104. 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.
105. 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.
106. 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.
107. 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.
108. 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.
109. 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.
110. 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.
111. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
112. 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.
113. 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.
114. 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.
115. 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.
116. 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.

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

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 16 October 2025 doi:10.20944/preprints202510.1295.v1

117. 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.
118. 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.