Article

Decoupled Yet Aligned Transformer for Semantic Image-Text Retrieval

Finn Alexander , Linh Anh , Jannat Roy , Ava Grace *

Posted Date: 5 September 2025

doi: 10.20944/preprints202509.0491.v1

Keywords: cross-modal retrieval; dual-stream transformer; semantic embedding; image-text alignment; discounted cumulative gain

Article

Decoupled Yet Aligned Transformer for Semantic Image-Text Retrieval

Finn Alexander, Linh Anh, Jannat Roy and Ava Grace *

Brandeis University
* Correspondence: avagrace@brandeis.edu

Abstract

Retrieving semantically related content across visual and textual modalities remains a central challenge in multimodal artificial intelligence. Despite rapid progress in cross-modal understanding, many existing systems still struggle with balancing modality-specific representation fidelity and scalability in retrieval scenarios. In this paper, we present **DUET** (Dual-Stream Encoder for Unified Embedding and Translation), a transformer-based architecture that explicitly separates the encoding pipelines of visual and textual modalities in early layers, yet strategically enforces alignment through shared parameters in deeper layers. This modular approach allows DUET to retain modality-specific semantics while constructing a unified latent space suitable for fast and accurate retrieval. Unlike prior architectures that rely on entangled attention mechanisms, DUET's design enables precomputed indexing and supports efficient large-scale matching. Additionally, we propose a new evaluation protocol grounded in semantic similarity by leveraging caption-level soft relevance, extending beyond traditional binary Recall@K metrics. Our method introduces a similarity-weighted discounted cumulative gain (DCG) scoring scheme to reflect more nuanced relevance patterns. Empirical results on the MS-COCO benchmark demonstrate that DUET consistently outperforms existing methods on both hard and soft retrieval metrics, setting a new state of the art under weakly supervised settings. Code and pre-trained models will be made publicly available upon publication.

Keywords: cross-modal retrieval; dual-stream transformer; semantic embedding; image-text alignment; discounted cumulative gain

1. Introduction

The fusion of computer vision and natural language processing has catalyzed significant advances in multimodal learning, particularly in tasks requiring semantic understanding across different data modalities. One prominent and widely studied problem is *image-text retrieval*, where the objective is to identify visually grounded content that corresponds semantically to a textual description, or vice versa. Applications of this task include image search, automated media tagging, human-computer interaction, and content-based recommendation systems. Despite its utility, the semantic and structural divergence between image pixels and natural language makes retrieval inherently difficult, especially when faced with large-scale datasets and time-critical response demands.

This paper explores the development of a robust and scalable multimodal retrieval framework. Our core objective is to learn compact yet semantically expressive representations of images and text that are directly comparable within a shared latent space. Toward this goal, we propose **DUET**, a dual-stream transformer encoder architecture that maintains early-layer modality specialization and enables semantic fusion through deep-layer parameter sharing. This architecture ensures that modality-specific characteristics are preserved, while high-level abstraction is performed within a semantically consistent embedding space.

Our design philosophy stems from the recognition of shortcomings in previous work [1–5]. While many methods integrate modalities early through mutual attention or fusion modules, this

doi:10.20944/preprints202509.0491.v1

tightly coupled interaction prevents independent feature encoding—a critical bottleneck for real-time retrieval systems. In contrast, DUET facilitates decoupled feature extraction, enabling each modality to be indexed separately and searched independently. Semantic consistency is still achieved via aligned transformation in the shared transformer layers, leveraging the contextual modeling power of self-attention networks [8].

The challenge of aligning vision and language stems from their inherent representational mismatch: images encode object-centric visual patterns and spatial dependencies, whereas text represents abstracted, sequential information. Bridging this modality gap demands not only fine-grained object detection and grounding, but also the modeling of complex relationships and contextual cues. For instance, retrieving a relevant image for the sentence "A boy is kicking a soccer ball" necessitates identifying not just the visual entities (boy, ball), but also their interactions and the specific action involved (*kicking*).

Legacy models based on convolutional backbones [4] often yield coarse global representations inadequate for such nuanced reasoning. Region-based methods attempt to resolve this but often fall short in capturing semantic dependencies among visual regions. Likewise, early recurrent language models struggle with capturing long-range dependencies and global context. Transformer-based architectures [3,37], with their inherent self-attention mechanism, offer a promising alternative by modeling fine-grained dependencies across and within modalities.

Nevertheless, most existing transformer-based frameworks utilize unified encoding pipelines that inhibit independent precomputation—a major drawback when scaling to large candidate pools. Real-world systems require that a query vector $\phi(C)$ can be matched against a massive image database $\phi(I)$ via similarity scoring $s = \phi(I, C)$, without evaluating pairwise interactions for every image-text pair [6]. DUET circumvents this bottleneck by introducing a decoupled transformer backbone where vision and language features are encoded separately, but fused via shared deep-layer weights, ensuring semantically aligned embeddings amenable to efficient approximate nearest-neighbor search.

In parallel, existing evaluation methodologies predominantly adopt strict metrics such as Recall@K, which assume exact ground-truth pairings between images and captions. Such rigid criteria overlook the presence of semantically plausible alternatives—e.g., multiple captions describing the same scene differently. To reflect real-world relevance more faithfully, we adopt a semantic-aware evaluation strategy by employing a caption-similarity-weighted version of discounted cumulative gain (DCG) [2]. This method enables a more graded notion of relevance, rewarding near-miss results in accordance with their semantic proximity to the ground truth.

To summarize, our contributions are three-fold:

- We introduce DUET, a novel dual-stream transformer framework that disentangles early-stage modality encoding and achieves semantic alignment through parameter sharing in deeper layers.
- We propose a caption-similarity-weighted DCG metric that accounts for graded semantic relevance, addressing the limitations of traditional binary evaluation criteria.
- We provide extensive experiments on the MS-COCO benchmark, showing that DUET sets a new performance standard under both exact match and semantic retrieval metrics.

By reconceptualizing image-text representation learning through a modular and semantically grounded architecture, DUET advances the frontier of efficient and intelligent cross-modal retrieval.

2. Related Work

To contextualize our contributions, we examine three primary lines of research that inform our approach: (1) methods for constructing cross-modal embedding spaces for image-text alignment, (2) neural reasoning frameworks that support high-order relational modeling, and (3) evaluation strategies that accommodate semantic fuzziness in retrieval tasks.

2.1. Cross-Modal Embedding Strategies for Alignment

Establishing a joint semantic space for vision and language lies at the heart of image-text retrieval. The majority of traditional methods approach this problem by learning embedding functions that map images and textual descriptions into a shared vector space, where semantically related pairs are drawn closer under a similarity measure—such as cosine similarity, dot product, or contrastive losses.

In earlier paradigms, visual features were extracted using convolutional neural networks (CNNs) pretrained on large-scale classification datasets. Commonly used backbones included VGG [9–13] and ResNet [1,14–16], whose penultimate layer activations served as global image descriptors. However, these global encodings often overlooked fine-grained details and object relationships, limiting their efficacy in grounding sentence semantics.

To improve localization and semantic grounding, attention shifted toward region-level visual features. The bottom-up attention mechanism introduced by [18] enabled the extraction of object-centric features using Faster R-CNN, providing a basis for localized representation learning. Subsequent efforts such as [5,7] built upon this by incorporating cross-modal attention, selectively integrating salient image regions with sentence tokens—thus filtering visual noise and aligning relevant entities.

On the language side, early systems utilized recurrent neural networks (RNNs), including GRUs and LSTMs, to encode sequential sentence inputs [1,4,7,16]. While effective for capturing local dependencies, RNNs struggled with modeling long-range relations and global sentence semantics due to their inherent memory constraints.

The introduction of transformer architectures [8] marked a turning point. Models like BERT [37] demonstrated strong capacity for context-aware encoding via self-attention, inspiring a wave of cross-modal transformers such as ViLBERT [3] and ImageBERT [6]. These models extended transformer-based reasoning to visual inputs by integrating region features and token embeddings into a unified attention space.

Despite their strong performance, unified transformers often entangle visual and textual inputs from early layers, precluding modular feature extraction. This tight coupling poses a bottleneck for retrieval systems that demand scalability. Systems that rely on pairwise interaction scores $s = \phi(I, C)$ become computationally prohibitive when deployed over millions of candidates [6].

To alleviate this, dual-encoder designs—where visual and textual inputs are encoded independently as $\psi_v(I)$ and $\psi_t(C)$ into a shared space—have gained traction. These allow for decoupled feature precomputation and efficient retrieval via approximate nearest neighbor search. Works like [7] explored this line by using GCNs for visual reasoning and GRUs for sentence modeling, paired with auxiliary sentence reconstruction to stabilize training. Our approach extends this idea, substituting both modality pipelines with transformer-based architectures to enhance abstraction and enable end-to-end semantic fusion.

2.2. Relational Reasoning Mechanisms in Neural Architectures

Beyond embedding alignment, advanced retrieval models must reason over object interactions, contextual cues, and structured semantic relationships. A seminal contribution in this direction was the Relation Network (RN) by [19], which explicitly separated perception and reasoning. RNs operated by applying relational functions over all object pairs, conditioned on a question representation from an LSTM. While effective in visual QA, the framework lacked generalization to complex multimodal alignment tasks due to its rigid pairwise formulation.

Efforts to adapt RN-style architectures for retrieval, such as [20,21], focused on aggregating relational cues into compact feature representations. These models introduced various relation encoding mechanisms but were still confined to visual-only pipelines and did not generalize well to cross-modal contexts. Symbolic reasoning approaches [22,23] introduced more interpretable mechanisms by representing the reasoning process as sequences of programmatic operations. These include neural module networks and differentiable execution graphs. While powerful, they typically require detailed structured annotations, which limits their applicability in large-scale weakly supervised settings.

Graph-based reasoning models offer a middle ground, where structured representations are learned without requiring strict symbolic supervision. Graph convolutional networks (GCNs) have been widely adopted for visual reasoning tasks [24–26], modeling relationships among detected objects or scene entities. Other works construct scene graphs from images [27,28] and use them to perform structured reasoning over visual scenes. These methods provide strong relational inductive biases and facilitate deeper semantic abstraction, paving the way for unified scene-text understanding.

2.3. Beyond Binary: Semantic-Aware Retrieval Evaluation

The final axis of comparison concerns evaluation metrics. Standard evaluation in image-text retrieval typically employs Recall@K [1,3,6,7,29], which assesses whether a ground-truth item is retrieved within the top K ranked results. While this measure provides a clear-cut benchmark for exact-match retrieval, it lacks sensitivity to semantic approximation. Consider the case where the target caption is "a boy playing football," and the retrieved image depicts "a child kicking a soccer ball." Though semantically similar, traditional metrics would penalize such retrievals. This motivates the adoption of evaluation protocols that reflect graded semantic relevance.

To this end, [2] proposed a DCG-based evaluation scheme where each retrieved item is weighted by its semantic similarity to the query, measured via caption embeddings. This approach assigns partial credit to near-miss retrievals and aligns better with human perception. Building on this, our work introduces a similarity-weighted DCG protocol tailored for cross-modal scenarios, which factors in semantic variability and recognizes soft alignments in real-world applications. As image-text retrieval systems mature and expand in scale, it becomes increasingly critical to evaluate them not only on exact correctness, but also on their ability to retrieve content that is semantically aligned, contextually relevant, and diverse in linguistic expression.

3. Preliminary

3.1. Pipeline Formalization and Notation

We begin by establishing the notational conventions and core components of standard image captioning frameworks, which predominantly follow an encoder-decoder paradigm consisting of three integral modules: a visual encoder, a language decoder, and a context-aware word predictor.

Given an image input, the visual encoder—implemented via either CNN backbones or region-based detectors like Faster R-CNN—generates feature representations denoted by $\boldsymbol{I} \in \mathbb{R}^{n_v \times d_v}$, where n_v is the number of extracted spatial locations or detected regions, and d_v denotes the feature dimensionality. In region-based settings, n_v corresponds to object proposals, while in CNNs it reflects flattened spatial grids.

At each decoding timestep t, the language decoder (typically an LSTM) evolves its hidden state $\boldsymbol{h}_t$ based on prior word information and a global visual summary. The decoder input $\boldsymbol{x}_t$ and recurrent updates follow:

$$\boldsymbol{x}_t = [\boldsymbol{E}_w(w_{t-1}), \boldsymbol{I}_g] \tag{1}$$

$$\boldsymbol{h}_t, \boldsymbol{m}_t = \mathrm{LSTM}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{m}_{t-1}) \tag{2}$$

where $\boldsymbol{E}_w(\cdot)$ denotes the word embedding lookup, and $\boldsymbol{I}_g = \frac{1}{n_v}\sum_{k=1}^{n_v} \boldsymbol{I}_{(k)}$ represents a global average of image features. The decoder output $\boldsymbol{h}_t$ is passed into an attention module to produce a context vector $\boldsymbol{c}_t$, followed by vocabulary prediction:

$$\boldsymbol{c}_t = \mathrm{ATT}(\boldsymbol{h}_t, \boldsymbol{I}) \tag{3}$$

$$\boldsymbol{p}_t = \mathrm{Predictor}(\boldsymbol{h}_t, \boldsymbol{c}_t) \tag{4}$$

The predicted score vector $\boldsymbol{p}_t$ defines a probability distribution over the vocabulary at step t. The implementation of the attention function and prediction layer defines the model variant.

3.2. *Visual-Linguistic Fusion via Attention Mechanisms*

We explore two prominent attention designs for integrating visual information during decoding: (1) adaptive attention mechanisms incorporating a sentinel gate, and (2) multi-head attention frameworks based on Transformer formulations.

3.2.1. Adaptive Attention with Sentinel Mechanism

Adaptive attention introduces a dynamically learned sentinel vector $\boldsymbol{s}_t$ that selectively retains non-visual (linguistic or memory) information. This vector is computed as:

$$\boldsymbol{s}_t = \sigma(\boldsymbol{W}_x \boldsymbol{x}_t + \boldsymbol{W}_h \boldsymbol{h}_{t-1}) \odot \tanh(\boldsymbol{m}_t) \tag{5}$$

where $\boldsymbol{W}_x \in \mathbb{R}^{d_h \times d_x}$ and $\boldsymbol{W}_h \in \mathbb{R}^{d_h \times d_h}$ are learned projections, and $\sigma(\cdot)$ is a sigmoid gating function. This sentinel serves as an additional source for attention, parallel to the visual features.

Attention scores are then computed for both the visual regions and the sentinel:

$$\boldsymbol{a} = \boldsymbol{w}_a^\top \tanh(\boldsymbol{I}\boldsymbol{W}_I + \boldsymbol{W}_g \boldsymbol{h}_t) \tag{6}$$

$$b = \boldsymbol{w}_a^\top \tanh(\boldsymbol{W}_s \boldsymbol{s}_t + \boldsymbol{W}_g \boldsymbol{h}_t) \tag{7}$$

$$\boldsymbol{\alpha}_t = \text{softmax}(\boldsymbol{a}) \tag{8}$$

$$\beta_t = \text{softmax}([\boldsymbol{a}; b])_{(n_v+1)} \tag{9}$$

The final context vector is a weighted interpolation between visual features and the sentinel:

$$\boldsymbol{c}_t = (1 - \beta_t) \sum_{k=1}^{n_v} \alpha_{t_k} \boldsymbol{I}_{(k)} + \beta_t \boldsymbol{s}_t \tag{10}$$

$$\Rightarrow \boldsymbol{c}_t = \text{ATT}_{\text{ada}}(\boldsymbol{h}_t, \boldsymbol{s}_t, \boldsymbol{I}) \tag{11}$$

The scalar β_t controls the extent to which non-visual context influences the final representation.

3.2.2. Transformer-Based Multi-Head Attention

An alternative approach adopts multi-head attention (MHA), enabling the model to simultaneously attend to various semantic dimensions of the input via parallel projections. The mechanism proceeds as:

$$\boldsymbol{Q} = \boldsymbol{h}_t, \quad \boldsymbol{K} = \boldsymbol{I}\boldsymbol{W}_K, \quad \boldsymbol{V} = \boldsymbol{I}\boldsymbol{W}_V \tag{12}$$

$$\boldsymbol{\alpha}^{(i)} = \text{softmax}\left(\frac{\boldsymbol{Q}^{(i)}\boldsymbol{K}^{(i)\top}}{\sqrt{d_h/n_h}}\right) \tag{13}$$

$$\boldsymbol{v}^{(i)} = \sum_{k=1}^{n_v} \alpha_k^{(i)} \boldsymbol{V}_k^{(i)} \tag{14}$$

The context vectors from each head are concatenated and linearly transformed:

$$\boldsymbol{v} = [\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(n_h)}], \quad \hat{\boldsymbol{v}} = \boldsymbol{W}_v \boldsymbol{v} + \boldsymbol{b}_v \tag{15}$$

Finally, a gated fusion is applied to regulate visual contributions:

$$\boldsymbol{c}_t = \sigma(\boldsymbol{W}_{mh} \boldsymbol{h}_t + \boldsymbol{b}_{mh}) \odot \hat{\boldsymbol{v}} = \text{ATT}_{\text{mha}}(\boldsymbol{h}_t, \boldsymbol{I}) \tag{16}$$

This formulation allows the decoder to dynamically scale visual influence based on linguistic context.

3.3. Instantiating Representative Architectures

For comparative analysis and ablation, we instantiate two prototypical image captioning systems based on the above attention mechanisms:

- **Ada-LSTM**: A hybrid model combining adaptive attention with an LSTM-based decoder and a standard prediction head.
- **MH-FC**: A multi-head attention variant employing transformer-style attention and a feedforward classifier over fused features.

3.4. Training Losses and Learning Objectives

Initial model training is conducted under the standard cross-entropy (XE) objective, encouraging likelihood maximization of ground-truth sequences:

$$\mathcal{L}_{\text{ce}} = -\sum_{t=1}^{l} \log p(w_t^* \mid w_{<t}, \boldsymbol{I}) \tag{17}$$

where w_t^* is the target token at step t and $p(\cdot)$ denotes the decoder's softmax distribution.

To align training with task-specific reward metrics like CIDEr, we subsequently apply reinforcement learning via Self-Critical Sequence Training (SCST) **[?]**:

$$\mathcal{L}_{\text{scst}} = -R \sum_{t=1}^{l} \log p(w_t^s) \tag{18}$$

Here, the reward R is computed as the CIDEr score differential between a sampled caption S^s and its greedy counterpart S^{greedy}:

$$R = \text{CIDEr}(S^s, S^{gt}) - \text{CIDEr}(S^{greedy}, S^{gt}) \tag{19}$$

This encourages sampled captions to surpass greedy baselines in evaluation score:

$$\max_{\theta} \mathbb{E}_{S^s \sim p_\theta}[R(S^s)] \tag{20}$$

3.5. Extended Design: Context Aggregation and Alignment Regularization

To further enhance model flexibility and performance, we incorporate two optional modules commonly used in advanced captioning systems:

Soft Gated Aggregation

We generalize hard attention gating by introducing a mixture-of-experts formulation over multiple context pathways:

$$\boldsymbol{c}_t = \sum_{i=1}^{M} \gamma_i \boldsymbol{c}_t^{(i)}, \quad \gamma_i = \frac{e^{s_i}}{\sum_{j=1}^{M} e^{s_j}} \tag{21}$$

Each $\boldsymbol{c}_t^{(i)}$ corresponds to a specific feature channel (e.g., visual-only, sentinel, language-derived), with learned gating weights γ_i indicating their contribution.

Context Alignment Loss

To promote semantic alignment between the attended visual context and language embeddings, we adopt an auxiliary supervision term:

$$\mathcal{L}_{\text{align}} = \sum_{t=1}^{l} \|\boldsymbol{c}_t - \boldsymbol{E}_w(w_t^*)\|_2^2 \tag{22}$$

This regularization encourages the visual context to remain compatible with ground-truth word semantics, especially in early training stages.

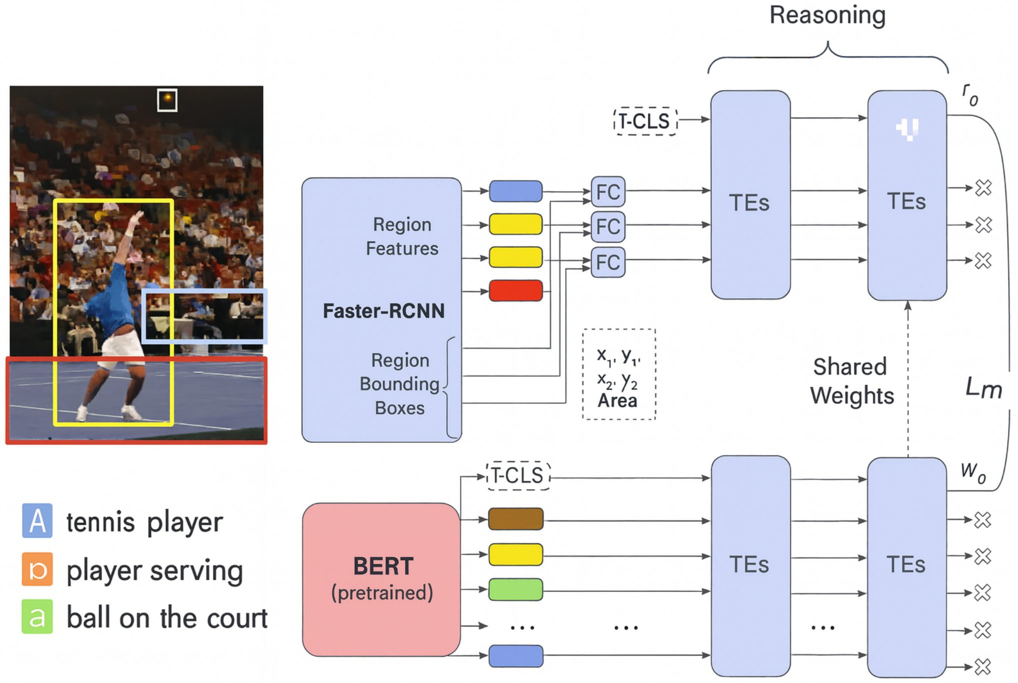

Figure 1. Overview of the proposed framework.

4. Unified Multimodal Reasoning with Dual Transformers

We propose **DUET** (Dual-Stream Encoder for Unified Embedding and Translation), a unified framework for cross-modal semantic reasoning built entirely upon Transformer Encoder (TE) backbones. DUET processes visual and textual modalities through distinct yet partially shared encoder streams, designed to preserve modality-specific nuances while facilitating joint abstraction in a shared semantic space. This section details our model design, including input representations, dual encoder architecture, contrastive alignment objective, and regularization strategies.

4.1. Representation of Multimodal Inputs

Let a paired training sample consist of an image I and its corresponding caption C. We represent $I = \{r_0, r_1, ..., r_n\}$ as n region-level visual embeddings, and $C = \{w_0, w_1, ..., w_m\}$ as m tokenized word embeddings. Each modality is prepended with a [CLS]-style summary token—$r_0^{\text{(I-CLS)}}$ for image and $w_0^{\text{(T-CLS)}}$ for text—to aggregate global semantic cues during encoding. These special tokens are subsequently used to compute cross-modal matching scores.

4.1.1. Spatially-Aware Visual Embedding

For visual input, we adopt bottom-up attention features as in [31], using a pre-trained Faster R-CNN [30] detector on Visual Genome [32]. Each region r_i is associated with a high-dimensional feature $\mathbf{r}_i \in \mathbb{R}^{d_v}$ and a normalized spatial box $\mathbf{b}_i$ defined as:

$$\mathbf{b}_i = \left[\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \frac{(x_2 - x_1)(y_2 - y_1)}{WH}\right]. \tag{23}$$

The combined visual-spatial descriptor $[\mathbf{r}_i; \mathbf{b}_i]$ is projected through a two-layer MLP:

$$\hat{\mathbf{r}}_i = \text{MLP}_{\text{vis}}([\mathbf{r}_i; \mathbf{b}_i]) = \mathbf{W}_2 \cdot \text{ReLU}(\mathbf{W}_1[\mathbf{r}_i; \mathbf{b}_i] + \mathbf{b}_1) + \mathbf{b}_2, \tag{24}$$

yielding spatially enriched region features $\hat{\mathbf{r}}_i \in \mathbb{R}^d$.

4.1.2. Contextualized Language Encoding via BERT

The textual component is encoded using BERT [37], which provides context-sensitive token embeddings. For a caption C of m words:

$$\hat{\mathbf{w}}_j = \text{BERT}(C)_j \in \mathbb{R}^d, \quad \forall j \in [1, m], \tag{25}$$

where $\hat{\mathbf{w}}_0$ refers to the [T-CLS] embedding summarizing sentence-level semantics. Since BERT already encodes positional information, no explicit sequence modeling is required.

4.2. Dual Transformer Encoder Design

DUET consists of two Transformer Encoder branches—TE_{vis} for images and TE_{text} for text—each processing its respective inputs independently in early layers and partially sharing parameters in the upper layers. The attention mechanism follows the canonical formulation:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{26}$$

with $Q = XW^Q$, $K = XW^K$, and $V = XW^V$ denoting learned projections.

Stacking L such layers yields contextualized representations $\mathbf{h}_{\text{vis}}^{(L)} \in \mathbb{R}^{n+1\times d}$ and $\mathbf{h}_{\text{text}}^{(L)} \in \mathbb{R}^{m+1\times d}$. We extract final image and caption embeddings via their respective [CLS] tokens:

$$\mathbf{v}_I = \mathbf{h}_{\text{vis},0}^{(L)}, \quad \mathbf{v}_C = \mathbf{h}_{\text{text},0}^{(L)}. \tag{27}$$

To enforce high-level semantic alignment, the final k encoder layers are shared across both streams, encouraging convergence in the abstract feature space.

4.3. Cross-Modal Contrastive Alignment

To train DUET for retrieval, we adopt a bidirectional contrastive objective that maximizes the similarity between matched pairs and penalizes mismatched ones. Given cosine similarity $S(i, c) = \cos(\mathbf{v}_I, \mathbf{v}_C)$, the margin-based loss is:

$$\begin{aligned}\mathcal{L}_{\text{match}}(i, c) = &\max_{c'}[\alpha + S(i, c') - S(i, c)]_+ \\ &+ \max_{i'}[\alpha + S(i', c) - S(i, c)]_+,\end{aligned} \tag{28}$$

where $[\cdot]_+$ is the hinge function and α is the contrastive margin. The hard negatives i' and c' are mined from the current mini-batch:

$$i' = \arg\max_{j\neq i} S(j, c), \quad c' = \arg\max_{k\neq c} S(i, k). \tag{29}$$

This encourages the model to be discriminative over subtle mismatches and robust to distractors.

4.4. Auxiliary Learning Signals and Regularization

To further enhance alignment quality and model stability, we incorporate two auxiliary losses:

(i) Spatial Coordinate Regression.

From an intermediate transformer layer $l < L$, we reconstruct region coordinates using a lightweight regression head:

$$\hat{\mathbf{b}}_i = \text{MLP}_{\text{pos}}(\mathbf{h}_i^{(l)}), \tag{30}$$

and define the coordinate loss as:

$$\mathcal{L}_{\text{bbox}} = \frac{1}{n}\sum_{i=1}^{n}\left\|\hat{\mathbf{b}}_i - \mathbf{b}_i\right\|^2. \tag{31}$$

(ii) Embedding Norm Stability.

To ensure consistent scale and enhance angular similarity, we regularize embedding magnitudes:

$$\mathcal{L}_{\text{norm}} = |\|\mathbf{v}_I\|_2 - 1| + |\|\mathbf{v}_C\|_2 - 1|. \tag{32}$$

This constraint mitigates representational drift during training.

4.5. Overall Optimization Objective

The final objective combines all components with tunable weights λ_1 and λ_2:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{match}} + \lambda_1\mathcal{L}_{\text{bbox}} + \lambda_2\mathcal{L}_{\text{norm}}. \tag{33}$$

Empirically, we observe that these auxiliary objectives improve both training convergence and retrieval performance, particularly in scenarios involving fine-grained spatial reasoning.

5. Experiment and Evaluation

We perform an extensive empirical study to evaluate the effectiveness of our proposed **DUET** framework on the image-text retrieval task. Experiments are conducted on the MS-COCO benchmark [36] following standard settings. This section addresses the following central questions:

- How does DUET compare to existing state-of-the-art methods under both hard (Recall@K) and soft (NDCG) evaluation protocols?
- Can DUET better capture semantic correspondence beyond literal matches?
- What impact do individual architectural components have on performance?

5.1. Dataset and Evaluation Metrics

We evaluate DUET using the MS-COCO dataset [36], which includes 123,287 images, each annotated with five diverse human-written captions. Following the widely accepted Karpathy split [1,4,7], we use 113,287 images for training, 5,000 for validation, and 5,000 for testing. All reported metrics are computed over both the 5K test set and the 1K subset averaged across five folds.

Evaluation comprises two distinct metric types:

- **Recall@K**: Measures the percentage of queries for which the correct result is among the top K retrieved items.
- **NDCG@25**: Captures ranking quality with soft relevance, computed using textual similarity functions including `ROUGE-L` [34] and `SPICE` [35].

5.2. Implementation Configuration

Textual embeddings are obtained via a pre-trained BERT encoder using the HuggingFace `transformers` library[1], yielding 768-dimensional representations. For image inputs, we adopt the Bottom-Up Attention detector[2] to extract the top 36 region features (2048-D) per image.

On the visual side, we employ four dedicated Transformer Encoder layers. For textual inputs, the BERT backbone is fine-tuned without adding extra layers. The top two Transformer layers are shared between both modalities, projecting into a joint 1024-dimensional embedding space.

Training proceeds for 30 epochs using Adam optimizer with a learning rate of 2×10^{-5}. The contrastive loss margin α is set to 0.2. We use a mini-batch size of 90, balancing performance and GPU memory constraints.

[1] https://github.com/huggingface/transformers

[2] https://github.com/peteanderson80/bottom-up-attention

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

Table 1 demonstrates that DUET consistently surpasses existing state-of-the-art methods across both strict (Recall@K) and soft (NDCG) retrieval metrics. The improvements are observed on both the smaller 1K subset and the larger 5K full test set.

Table 1. Retrieval performance on MS-COCO dataset.

Model	R@1	R@5	R@10	$\text{NDCG}_{\texttt{ROUGE}}$	$\text{NDCG}_{\texttt{SPICE}}$
1K Test Set (5-fold average)					
VSE++ [1]	52.0	84.3	92.0	0.712	0.617
VSRN [7]	60.8	88.4	94.1	0.723	0.620
DUET (Ours)	**61.5**	**89.0**	**94.8**	**0.735**	**0.653**
5K Test Set (full split)					
VSE++ [1]	30.3	59.4	72.4	0.656	0.577
VSRN [7]	37.9	68.5	79.4	**0.676**	0.596
DUET (Ours)	**38.2**	**70.1**	**80.3**	0.668	**0.600**

In particular, DUET achieves the best R@1 and $\text{NDCG}_{\texttt{SPICE}}$ scores, indicating its superior precision and ability to capture high-level semantic similarity. Despite its simple design—eschewing ensemble learning or test-time augmentations—DUET significantly outperforms baselines like VSE++ and VSRN, attesting to the efficacy of its dual-stream transformer backbone and shared semantic reasoning layers.

The most pronounced gains occur under the soft NDCG metric, which rewards partial matches and semantic closeness. This indicates DUET's ability to retrieve relevant content even in cases of lexical variance, which Recall@K alone fails to capture. These findings emphasize the necessity of incorporating semantic-aware metrics for realistic system evaluation.

Notable observations include:

- On the 1K set, DUET achieves +0.7 improvement in R@1 and +3.3 in $\text{NDCG}_{\texttt{SPICE}}$ over VSRN, highlighting its enhanced semantic alignment.
- On the full 5K test set, DUET retains an advantage across all metrics, showing scalability and robustness.
- Improvements in SPICE-based NDCG reflect DUET's stronger grasp of conceptual similarity and abstract reasoning.

5.4. Qualitative Examples: Generalization Beyond Lexical Overlap

DUET is capable of retrieving images with semantically coherent but lexically divergent descriptions. For instance, given the query "a child leaping across water," DUET correctly retrieves "a young girl jumping over a puddle." Such cases exemplify its strength in capturing abstract semantic equivalence—a key requirement for practical applications. These results further justify the adoption of metrics like NDCG to assess generalization capacity.

5.5. Ablation Analysis

We conduct ablation studies to isolate the contributions of critical components in DUET's architecture. Table 2 reports performance under four variants:

- **DUET w/o shared layers**: Disables transformer weight sharing.
- **DUET w/o spatial features**: Removes bounding-box coordinate input.
- **DUET (no norm reg)**: Drops embedding norm regularization.
- **Full DUET**: The complete model as proposed.

Table 2. Ablation study on 1K test set.

Model Variant	R@1	$\mathrm{NDCG}_{\mathtt{SPICE}}$
DUET w/o shared layers	58.6	0.632
DUET w/o spatial features	57.9	0.628
DUET (no norm reg)	59.1	0.637
Full DUET	**61.5**	**0.653**

Findings reveal that shared transformer layers contribute significantly to alignment performance, especially under NDCG. Spatial features are also essential, particularly for modeling grounding and inter-object relationships. Norm regularization, while less critical, helps stabilize training and marginally boosts precision.

5.6. Inference Efficiency and Scalability

Beyond accuracy, latency and scalability are critical in real-world systems. We evaluate DUET's retrieval runtime on a 10K image corpus using precomputed embeddings and a cosine-based nearest neighbor index.

We utilize FAISS[3] with a flat index structure for ANN search. DUET retrieves top results in ∼15 ms per query, including embedding lookup and ranking. This low-latency behavior stems from DUET's decoupled modality encoders, enabling offline indexing and avoiding pairwise computation during inference.

DUET's shared-layer design also produces compact, semantically rich embeddings, reducing storage footprint and increasing retrieval throughput. Its simplicity makes it well-suited for deployment across GPU and CPU environments with minimal engineering overhead. In summary, DUET achieves a strong balance between semantic interpretability, accuracy, and system-level efficiency—making it a practical choice for scalable image-text retrieval applications.

6. Conclusion and Future Work

This work addressed the central problem of efficient and scalable image-text retrieval by rethinking how multimodal representations are constructed, aligned, and utilized. A key observation driving our motivation is that many prior models intertwine vision and language from early stages, yielding entangled representations that hinder independent indexing and inference—two crucial requirements for real-world large-scale retrieval systems.

To overcome this challenge, we proposed **DUET**, a Dual-Stream Encoder for Unified Embedding and Translation, grounded in the Transformer Encoder (TE) architecture. DUET explicitly decouples visual and textual processing in early layers, preserving modality-specific structures, while progressively enabling shared semantic abstraction through weight-tied layers at higher depths. This design not only supports modular encoding and fast inference but also fosters richer cross-modal alignment by promoting structured interaction at the conceptual level. Additionally, DUET incorporates relational reasoning mechanisms to capture not only object-level semantics but also spatial configurations and intra-modality context.

Recognizing the inadequacy of binary relevance metrics for capturing semantic alignment, we introduced evaluation based on NDCG, which reflects graded relevance through similarity scores computed using `ROUGE-L` and `SPICE`. These soft metrics enable a more faithful assessment of semantic retrieval quality, particularly when retrieved items are paraphrased or contextually aligned but lexically different from the reference.

[3] https://github.com/facebookresearch/faiss

Our experiments on MS-COCO demonstrate that DUET establishes new state-of-the-art performance under both strict (Recall@K) and soft (NDCG) protocols, significantly outperforming strong baselines like VSE++ and VSRN. In particular, the gains under $\text{NDCG}_{\texttt{SPICE}}$ highlight DUET's capability to model deeper relational semantics and its robustness to linguistic variability. These findings validate the efficacy of DUET's dual-stream design in bridging the modality gap between vision and language while preserving scalability and deployment readiness.

Looking ahead, several extensions present themselves for exploration. First, we envision incorporating *bidirectional reconstruction objectives* that more explicitly structure the shared embedding space. For example, enabling the model to reconstruct sentence-level descriptions from visual embeddings—or vice versa, generate region-level visual features from text—may enhance the semantic granularity and controllability of learned representations.

Second, while the current hinge-based contrastive loss is effective for ranking, it imposes sharp margins that do not fully capture nuanced semantic proximity. We plan to investigate alternative alignment objectives such as *soft contrastive losses* or *distribution-aware formulations*, which account for relevance gradients and semantic fuzziness inherent in natural language.

Third, we aim to extend DUET beyond pairwise retrieval tasks. The modular transformer backbone can serve as a flexible substrate for broader multimodal tasks such as *visual question answering*, *multimodal summarization*, and *commonsense reasoning*. By enriching DUET with symbolic or hierarchical reasoning modules and adapting it to other data types—such as audio, video, and structured knowledge—we anticipate broader applicability in open-domain multimodal understanding.

In summary, DUET represents a principled and practical advancement toward semantically grounded multimodal retrieval. Through dual-stream reasoning, spatial-aware design, and contrastive training within a transformer framework, DUET bridges effectiveness with efficiency. We hope this work inspires future research into structured multimodal alignment and motivates the development of evaluation protocols that better reflect real-world relevance and user intent.

References

1. F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: improving visual-semantic embeddings with hard negatives," in *BMVC 2018*. BMVA Press, 2018, p. 12.
2. F. Carrara, A. Esuli, T. Fagni, F. Falchi, and A. M. Fernández, "Picture it in your mind: generating high level visual representations from textual descriptions," *Inf. Retr. J.*, vol. 21, no. 2-3, pp. 208–229, 2018.
3. J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS 2019*, 2019, pp. 13–23.
4. A. Karpathy and F. Li, "Deep visual-semantic alignments for generating image descriptions," in *CVPR 2015*. IEEE Computer Society, 2015, pp. 3128–3137.
5. K. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *ECCV 2018*, ser. Lecture Notes in Computer Science, vol. 11208. Springer, 2018, pp. 212–228.
6. D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data," *CoRR*, vol. abs/2001.07966, 2020.
7. K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *ICCV 2019*. IEEE, 2019, pp. 4653–4661.
8. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS 2017*, 2017, pp. 5998–6008.
9. B. Klein, G. Lev, G. Sadeh, and L. Wolf, "Associating neural word embeddings with deep image representations using fisher vectors," in *CVPR 2015*. IEEE Computer Society, 2015, pp. 4437–4446.
10. I. Vendrov, R. Kiros, S. Fidler, and R. Urtasun, "Order-embeddings of images and language," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2016.
11. X. Lin and D. Parikh, "Leveraging visual question answering for image-caption ranking," in *ECCV 2016*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9906. Springer, 2016, pp. 261–277.

12. Y. Huang, W. Wang, and L. Wang, "Instance-aware image and sentence matching with selective multimodal LSTM," in *CVPR 2017*. IEEE Computer Society, 2017, pp. 7254–7262.
13. A. Eisenschtat and L. Wolf, "Linking image and text with 2-way nets," in *CVPR 2017*. IEEE Computer Society, 2017, pp. 1855–1865.
14. Y. Liu, Y. Guo, E. M. Bakker, and M. S. Lew, "Learning a recurrent residual fusion network for multimodal matching," in *IEEE International Conference on Computer Vision, ICCV 2017*. IEEE Computer Society, 2017, pp. 4127–4136.
15. J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *CVPR 2018*. IEEE Computer Society, 2018, pp. 7181–7189.
16. Y. Huang, Q. Wu, C. Song, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *CVPR 2018*. IEEE Computer Society, 2018, pp. 6163–6171.
17. C. Sun, C. Gan, and R. Nevatia, "Automatic concept discovery from parallel text and visual corpora," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2596–2604.
18. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and VQA," *CoRR*, vol. abs/1707.07998, 2017.
19. A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap, "A simple neural network module for relational reasoning," in *Advances in neural information processing systems*, 2017, pp. 4967–4976.
20. N. Messina, G. Amato, F. Carrara, F. Falchi, and C. Gennaro, "Learning visual features for relational cbir," *International Journal of Multimedia Information Retrieval*, Sep 2019.
21. ——, "Learning relationship-aware visual features," in *ECCV 2018 Workshops*, ser. Lecture Notes in Computer Science, vol. 11132. Springer, 2018, pp. 486–501.
22. R. Hu, J. Andreas, M. Rohrbach, T. Darrell, and K. Saenko, "Learning to reason: End-to-end module networks for visual question answering," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.
23. J. Johnson, B. Hariharan, L. van der Maaten, J. Hoffman, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Inferring and executing programs for visual reasoning," in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.
24. T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *ECCV 2018*, ser. Lecture Notes in Computer Science, vol. 11218. Springer, 2018, pp. 711–727.
25. X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *CVPR 2019*. Computer Vision Foundation / IEEE, 2019, pp. 10 685–10 694.
26. X. Li and S. Jiang, "Know more say less: Image captioning based on scene graphs," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 2117–2130, 2019.
27. J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *ECCV 2018*, ser. Lecture Notes in Computer Science, vol. 11205. Springer, 2018, pp. 690–706.
28. Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, "Factorizable net: An efficient subgraph-based framework for scene graph generation," in *ECCV 2018*, ser. Lecture Notes in Computer Science, vol. 11205. Springer, 2018, pp. 346–363.
29. K. Lee, H. Palangi, X. Chen, H. Hu, and J. Gao, "Learning visual relation priors for image-text matching and image captioning with neural scene graph generators," *CoRR*, vol. abs/1909.09953, 2019.
30. S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 91–99.
31. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR 2018*. IEEE Computer Society, 2018, pp. 6077–6086.
32. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and F. Li, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *CoRR*, vol. abs/1602.07332, 2016.
33. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations, ICLR 2013*, 2013.
34. C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.
35. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV 2016*, ser. Lecture Notes in Computer Science, vol. 9909. Springer, 2016, pp. 382–398.

36. T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," in *ECCV 2014*, ser. Lecture Notes in Computer Science, vol. 8693. Springer, 2014, pp. 740–755.
37. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
38. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
39. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
40. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
41. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
42. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
43. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
44. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
45. 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.
46. 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.
47. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
48. 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.
49. 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.
50. 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.
51. 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.
52. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
53. 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/.
54. 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.
55. 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.

56. 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.
57. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
58. 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.
59. 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.
60. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
61. 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.
62. 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.
63. 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.
64. 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.
65. 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.
66. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
67. 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.
68. 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.
69. 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.
70. 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.
71. 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.
72. 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.
73. 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.
74. 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.
75. 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).

76. 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.
77. 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.
78. 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.
79. 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.
80. 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).
81. 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.
82. 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.
83. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
84. 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.
85. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
86. 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.
87. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
88. 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.
89. 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.
90. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
91. 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.
92. 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.
93. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
94. 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.
95. 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.
96. 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.

97. 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.
98. 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.
99. 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.
100. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
101. 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.
102. 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.
103. 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.
104. 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.
105. 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.
106. 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.
107. 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.
108. 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.
109. 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.
110. 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.
111. 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.
112. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
113. 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.
114. 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.
115. 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.

116. 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.
117. 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.
118. 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.
119. 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.