Anticipatory Semantics with Bidirectional Guidance for Image Captioning

Noémie Laurent *, Elodie Fairchild, Arthur Delvaux

Posted Date: 17 September 2025

doi: 10.20944/preprints202509.1497.v1

Keywords: image captioning; anticipatory semantics; bidirectional attention; reinforcement optimization; dual-pass refinement; vision-language generation

 doi:10.20944/preprints202509.1497.v1

Article

Anticipatory Semantics with Bidirectional Guidance for Image Captioning

Noémie Laurent *, Elodie Fairchild and Arthur Delvaux

Université libre de Bruxelles, Brussels, Belgium
* Correspondence: noemielaurent@ulb.be

Abstract

Producing captions that are not only grammatically fluent but also semantically faithful to visual content has long stood as a central problem at the junction of computer vision and natural language processing. Conventional encoder-decoder frameworks with attention modules, although powerful, typically confine the decoding process to a retrospective scope: every prediction is conditioned solely on historical tokens, thereby ignoring what future semantics may dictate. This retrospective bias hinders models from fully capturing scene-level coherence. To address this limitation, we introduce **FISRA** (Future-Infused Semantic Revision Architecture), a dual-pass attention paradigm that supplements training with anticipatory semantic signals. Specifically, FISRA first constructs a global caption hypothesis serving as a semantic scaffold and then refines the decoding trajectory through revision-oriented attention that aligns each step with both preceding context and projected continuations. This design enhances coherence by weaving forward-looking cues into the generative process. The framework is model-agnostic and integrates seamlessly with existing attention-based captioners. Empirical evaluations on MS-COCO reveal that FISRA consistently advances the state of the art, delivering 133.4 CIDEr-D on the Karpathy split and 131.6 CIDEr-D on the official evaluation server. These findings confirm that our approach significantly strengthens semantic alignment and alleviates the exposure bias endemic to unidirectional captioning systems.

Keywords: image captioning; anticipatory semantics; bidirectional attention; reinforcement optimization; dual-pass refinement; vision-language generation

1. Introduction

The automatic generation of natural language captions for images remains a prominent challenge in artificial intelligence, demanding not only an accurate grasp of the visual scene but also the ability to articulate it in coherent textual form. This problem sits at the convergence of computer vision and natural language processing and underpins a range of applications including multimodal retrieval [14,33,34], assistive interfaces, and interactive human-machine communication [8,30,38]. Ultimately, the task aspires to design models capable of narrating complex environments with human-level expressiveness.

Existing methodologies are dominated by encoder-decoder pipelines [32]. Typically, convolutional neural networks (CNNs) encode spatially rich features from an image, and recurrent units such as LSTMs or GRUs decode them into sentence-level descriptions. The adoption of attention mechanisms, inspired by neural machine translation [3], has allowed these models to highlight salient spatial regions during generation [39]. Yet, the dominant training paradigm, maximum likelihood estimation with "Teacher Forcing" [37], creates a disconnect between training and inference—commonly referred to as exposure bias [28]. During training, ground-truth words condition the decoder, while at inference, the system must rely solely on its prior predictions, compounding any errors that emerge.

To mitigate this discrepancy, reinforcement learning (RL) approaches such as REINFORCE [36] have been explored to directly optimize evaluation metrics like CIDEr and SPICE. However, these

 doi:10.20944/preprints202509.1497.v1

RL-based strategies still conform to a strictly left-to-right generative order, leaving them blind to the semantic content yet to come. In effect, they fail to exploit global awareness of the caption's intended trajectory.

This unidirectional constraint is a fundamental weakness. Predictions at each time step are tethered exclusively to the past, precluding access to semantic signals embedded in subsequent tokens. Consider a caption fragment "A boy is"—conventional decoders may infer "standing" as a plausible continuation. Yet the true caption "A boy is playing baseball" demands recognition of the downstream concept "playing." Absent this forward-looking capacity, models risk settling for shallow heuristics rather than authentic scene interpretation.

To overcome these shortcomings, we propose **FISRA**, a semantic revision framework that enforces bidirectional reasoning through a two-phase decoding process. Initially, a draft caption is generated as a global hypothesis, serving as a proxy for future semantics. Subsequently, an auxiliary refinement module revisits the sequence with masked word prediction guided by semantic attention that draws simultaneously from preceding and succeeding tokens. This mechanism encourages the model to align its predictions with broader scene semantics rather than merely local cues.

During inference, FISRA operates in two stages: first, it generates a preliminary caption via greedy decoding, which acts as a semantic blueprint. Second, this caption informs a refined beam-search decoding pass, wherein the auxiliary module supplies contextual cues that adjust local predictions to fit the global narrative. This dual-pass pipeline strikes a balance between computational feasibility and semantic fidelity.

We further validate FISRA by embedding it into widely adopted captioning baselines such as Att2all [29], Up-Down [2], and AoANet [13]. Across the MS-COCO benchmark, our augmented models consistently surpass their baselines under standard metrics. For instance, coupling FISRA with the Up-Down model lifts CIDEr-D from 123.4 to 128.8 on the Karpathy split, evidencing the robustness of anticipatory semantic guidance.

To summarize, our contributions are as follows:

- We present FISRA, a dual-pass captioning framework that injects global semantic anticipation into the decoding process, enabling predictions that are both contextually grounded and forward-aware.
- Our design is modular and easily applicable to attention-based models trained with reinforcement learning, consistently improving captioning quality.
- Extensive experimentation on MS-COCO confirms that FISRA delivers substantial performance gains, establishing new benchmarks over strong existing baselines.

In essence, this study underscores the importance of predictive semantics in vision-language generation and paves the way for future research into bidirectional reasoning paradigms for sequence modeling.

2. Related Work

Research in automatic image captioning has undergone rapid transformation in recent years, largely driven by the progress of deep neural networks and multimodal representation learning. A long-standing taxonomy separates methods into bottom-up strategies [7,9,16,18,25], which focus on local object-centric features and region-level proposals, and top-down strategies [24,32,35,41,44], which emphasize global semantics and contextual cues. The present study is more closely aligned with the latter, following the encoder-decoder paradigm [32] in which convolutional or transformer-based encoders extract holistic visual representations that are subsequently translated into natural language sequences by recurrent or attention-driven decoders.

Within this broader framework, two methodological threads have been especially influential: the development of attention modules and the adoption of reinforcement learning objectives. Attention architectures guide captioners toward semantically salient image content, whereas reinforcement learning addresses the mismatch between training likelihood objectives and evaluation-time metrics.

 doi:10.20944/preprints202509.1497.v1

A more recent line of inquiry further considers the role of bidirectional or future-aware reasoning as an avenue for bridging local consistency with global semantic fidelity. We review these research strands in detail below.

2.1. Attention Mechanisms in Image Captioning

The introduction of attention mechanisms has profoundly shaped modern captioning models by enabling adaptive focus on informative visual regions. Originally inspired by advances in neural machine translation and object recognition, attention modules provide a dynamic means to align image features with textual tokens, thereby enhancing semantic grounding and improving descriptive quality.

The seminal work of Xu et al. [39] introduced both soft and hard attention for captioning tasks. Soft attention computes differentiable weighted sums over spatial features, while hard attention samples regions stochastically, yielding greater interpretability but higher training variance. Building upon this, Anderson et al. [2] proposed bottom-up attention that integrates region-level proposals from Faster R-CNN trained on Visual Genome [15], leading to sharper object localization and improved caption quality. Guo et al. [11] extended this approach with a two-stage decoder that refines coarse initial sentences into more coherent outputs.

Subsequent work explored richer structural modeling. Yao et al. [42] incorporated explicit relational graphs encoding semantic and spatial dependencies among objects, while Yu et al. [27] advanced predictive capacity by proposing a forward-looking mechanism that simultaneously forecasts multiple tokens. Herdade et al. [12] recast captioning within a transformer architecture augmented by object relations, and Li et al. [17] introduced EnTangled Attention (ETA), designed to fuse multimodal signals through jointly learned semantic and visual dependencies. Other hierarchical strategies include multi-level attention [43], as well as the Adaptive Object Attention (AoA) model of Huang et al. [13], which adaptively modulates relevance by conditioning attention scores on the decoder's hidden state.

In summary, the trajectory of attention research has progressed from simple soft-focus modules toward complex, multi-level relational architectures, progressively enhancing the ability of captioning systems to capture fine-grained scene semantics and generate more fluent descriptions.

2.2. Reinforcement Learning for Optimizing Caption Quality

Despite the success of attention mechanisms, models trained purely with maximum likelihood estimation (MLE) exhibit a discrepancy between training and inference conditions, known as exposure bias [28]. Reinforcement learning (RL) has thus been introduced as a corrective paradigm, aligning model training with evaluation metrics such as BLEU [26], ROUGE [19], METEOR [4], CIDEr [31], and SPICE [1], all of which are non-differentiable under standard supervision.

Liu et al. [22] proposed SPIDEr, a reward function that harmonizes SPICE and CIDEr to more closely reflect human judgment. Their approach leverages Monte Carlo rollouts to approximate long-term rewards. Subsequent extensions, such as the temporal difference learning scheme [5], assign action-level credit based on delayed impact, thereby refining policy optimization. Gao et al. [10] enhanced sample efficiency through n-step advantage estimation and hybrid rollout strategies. Among these innovations, self-critical sequence training (SCST) [29] stands out as a particularly impactful framework, introducing greedy decoding outputs as a reward baseline. SCST effectively reduces gradient variance while providing stable optimization, establishing itself as the standard approach for RL-based captioning.

Through these contributions, RL has broadened the optimization landscape by explicitly balancing fluency with semantic alignment. Nevertheless, challenges persist in terms of reward design, policy stability, and the extent to which automatic metrics capture human perceptual quality.

2.3. Future-Aware and Bidirectional Reasoning in Generation

Although attention and RL significantly advance caption quality, most systems still generate language autoregressively from left to right. This strictly unidirectional design prevents the decoder from

 doi:10.20944/preprints202509.1497.v1

exploiting cues that appear later in the sentence, leading to incomplete or myopic semantic reasoning. For instance, when generating the phrase "A boy is", a left-to-right decoder may prematurely settle on "standing", neglecting that the ground truth continuation "playing baseball" conveys a richer and more precise meaning.

To address this deficiency, several studies have proposed architectures capable of forward-looking or bidirectional reasoning. Approaches that generate preliminary full captions and subsequently refine them through revision decoders have shown promise in enhancing coherence. Other designs leverage semantic masking and reconstruction objectives to jointly learn from both preceding and subsequent tokens. These mechanisms resonate with ideas from masked language modeling and non-autoregressive generation, highlighting the potential of integrating global context into captioning systems. Although not yet pervasive in mainstream benchmarks, such methods chart an important research trajectory for future-aware captioning.

2.4. Summary and Positioning

In conclusion, the literature on captioning research reveals three dominant streams: attention architectures that provide dynamic grounding, reinforcement learning frameworks that align training with evaluation metrics, and nascent bidirectional reasoning mechanisms that incorporate anticipatory semantics. Our work synthesizes these directions by situating within the top-down captioning paradigm, employing attention and reinforcement optimization, and introducing a dual-pass refinement strategy that explicitly integrates both past and future cues. Through this lens, we aim to close the gap between locally fluent captioning and globally coherent, semantically robust generation.

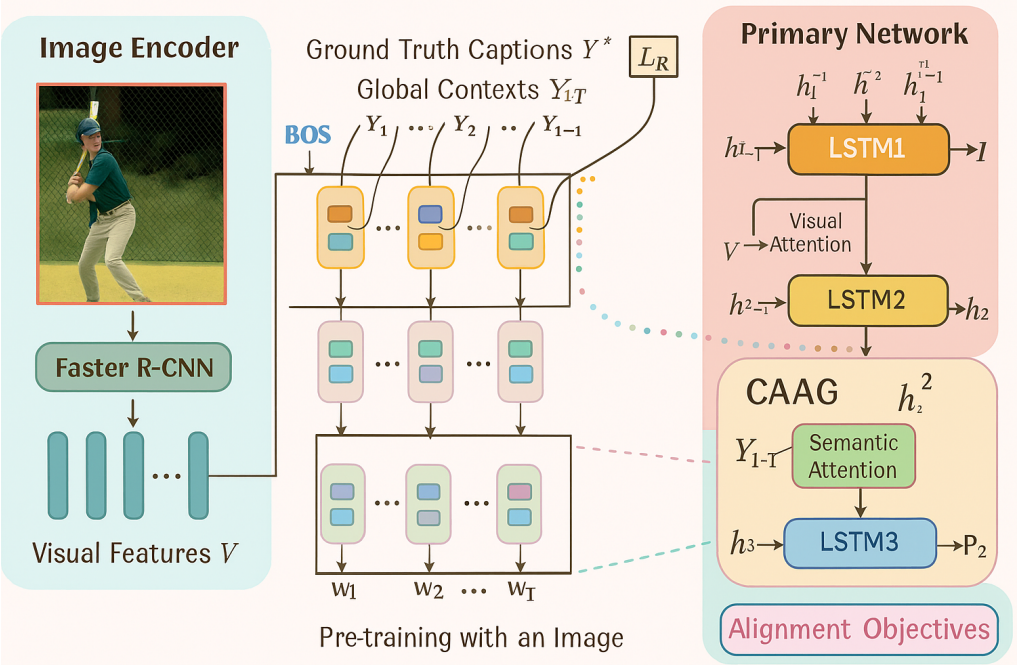

Figure 1. The schematic overview representation of the entire framework.

3. Methodological Framework: Dual-Path Reasoning for Image Captioning

We introduce **DUPLEX** (Dual-path Unified Planning and EXecution network), a novel framework specifically designed to enhance semantic consistency in image captioning by combining traditional autoregressive decoding with global semantic refinement. Unlike conventional encoder-decoder pipelines that rely solely on left-to-right token generation, DUPLEX explicitly leverages both past and future semantics, enabling captions that are not only grammatically fluent but also globally coherent. This section provides a detailed description of the model design, organized into multiple components and accompanied by discussions on its computational properties and interpretability.

 doi:10.20944/preprints202509.1497.v1

3.1. Overview of DUPLEX Framework

The DUPLEX framework builds upon the encoder-decoder paradigm but enriches it with a dual-path reasoning strategy. One path functions as a standard autoregressive decoder that generates tokens sequentially, ensuring fluent sentence construction. In parallel, a second path performs semantic refinement by analyzing the full caption holistically and adjusting token predictions accordingly. During training, the two paths are optimized jointly with complementary objectives, while during inference their outputs are fused into a final distribution. The design principle is to provide models with the capacity to reason over both local (short-term) and global (long-term) dependencies, thereby reducing the semantic drift common in unidirectional captioners.

3.2. Region-Level Visual Encoding

To ground the model in image semantics, we extract a set of region-level embeddings from the input image I using a Faster R-CNN detector trained on Visual Genome [15]. This process yields features $\boldsymbol{V} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_k\}$ where each $\boldsymbol{v}_i \in \mathbb{R}^{d_I}$ encodes local semantic and spatial information. The number of proposals k adapts to the visual complexity of the image, typically ranging from a few dozen in simple scenes to several hundred in cluttered environments. These embeddings are complemented by a global descriptor $\bar{\boldsymbol{v}} = \frac{1}{k}\sum_{i=1}^{k} \boldsymbol{v}_i$, which captures holistic scene semantics. Such a dual representation (region-level + global) allows DUPLEX to flexibly balance fine-grained detail and coarse contextual information.

3.3. Primary Autoregressive Path

The first decoding stream of DUPLEX is instantiated using the Up-Down architecture [2]. It consists of two stacked LSTMs, each with a distinct role. LSTM_1 operates as an attention controller: at timestep t, it takes as input the concatenation of the previous word embedding $\boldsymbol{x}_t$, the hidden state $\boldsymbol{h}_{t-1}^2$ from LSTM_2, and the global visual descriptor $\bar{\boldsymbol{v}}$. Its hidden state $\boldsymbol{h}_t^1$ is used to compute attention weights over $\boldsymbol{V}$, producing an attended feature vector $\hat{\boldsymbol{v}}_t$. LSTM_2 then fuses $\hat{\boldsymbol{v}}_t$ with $\boldsymbol{h}_t^1$ to update the language state $\boldsymbol{h}_t^2$, which is projected into vocabulary space to yield the word distribution p_t^1. This autoregressive mechanism provides fluent word-by-word caption generation but is restricted by its unidirectional design.

3.4. Auxiliary Semantic Refinement Path

The auxiliary path supplements the primary decoder by introducing bidirectional semantic reasoning. Instead of generating tokens autoregressively, it takes the full caption $Y_{1:T}$ produced by the primary path and embeds it as $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$. A masked attention mechanism computes contextual vectors $\boldsymbol{c}_t$ for each timestep by attending over $\boldsymbol{X} \setminus \{\boldsymbol{x}_t\}$. This ensures that predictions are based on both past and future tokens, similar to masked language modeling objectives used in pretraining. The resulting $\boldsymbol{c}_t$ is concatenated with the autoregressive hidden state $\boldsymbol{h}_t^2$ and passed into a refinement LSTM (LSTM_3), producing $\boldsymbol{h}_t^3$. A secondary softmax layer outputs the auxiliary distribution p_t^2, which emphasizes global consistency and long-range semantics.

3.5. Cross-Path Interaction Mechanisms

A central challenge in dual-path systems is ensuring that the two streams complement rather than contradict one another. DUPLEX introduces explicit interaction layers that allow signals from the auxiliary path to guide the autoregressive decoder during training. For instance, $\boldsymbol{h}_t^3$ can be back-projected into the hidden space of LSTM_2, serving as a corrective feedback signal. Similarly, agreement between p_t^1 and p_t^2 is monitored to regulate confidence, encouraging the model to rely more on auxiliary predictions when autoregressive uncertainty is high. This design not only stabilizes optimization but also fosters deeper semantic alignment between the two paths.

 doi:10.20944/preprints202509.1497.v1

Algorithm 1: Training Procedure of DUPLEX

Input: Image-caption dataset $\mathcal{D} = \{(I_i, Y_i^*)\}_{i=1}^N$; Pretrained Faster R-CNN detector ϕ_{rcnn}
Word embedding matrix $\mathcal{E}$, vocabulary $\mathcal{V}$; Hyperparameters $\alpha, \lambda_1, \lambda_2$
Output: Trained captioning model parameters (θ_p, θ_a)

Initialize primary decoder θ_p, auxiliary decoder θ_a ;
foreach *epoch* **do**
  foreach $(I, Y^*) \in \mathcal{D}$ **do**
    `// Step 1: Visual Feature Extraction`
    Extract region features $\boldsymbol{V} = \phi_{\text{rcnn}}(I)$;
    Compute global visual mean $\bar{\boldsymbol{v}} = \frac{1}{k}\sum_i \boldsymbol{v}_i$;
    `// Step 2: Primary Caption Generation`
    for $t = 1$ **to** T **do**
      Get previous token embedding $\boldsymbol{x}_t = \mathcal{E}(y_t)$;
      Compute input to LSTM_1: $\boldsymbol{z}_t^1 = [\boldsymbol{x}_t; \boldsymbol{h}_{t-1}^2; \bar{\boldsymbol{v}}]$;
      Update LSTM_1: $\boldsymbol{h}_t^1 = \text{LSTM}_1(\boldsymbol{z}_t^1)$;
      Compute attention weights $\alpha_{i,t}$; feature $\hat{\boldsymbol{v}}_t = \sum_i \alpha_{i,t}\boldsymbol{v}_i$;
      Compute input to LSTM_2: $\boldsymbol{z}_t^2 = [\hat{\boldsymbol{v}}_t; \boldsymbol{h}_t^1]$;
      Update LSTM_2: $\boldsymbol{h}_t^2 = \text{LSTM}_2(\boldsymbol{z}_t^2)$;
      Compute primary distribution $p_t^1 = \text{softmax}(W_{out}\boldsymbol{h}_t^2 + b_{out})$;
    end
    Sample sentence $\hat{Y} = \{y_1, y_2, \ldots, y_T\}$ from p_t^1 (greedy or sampling) ;
    `// Step 3: Auxiliary Contextual Refinement`
    Obtain embeddings $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ for $\hat{Y}$;
    for $t = 1$ **to** T **do**
      Compute attention weights $\beta_{i,t}$; context vector $\boldsymbol{c}_t = \sum_i \beta_{i,t}\boldsymbol{x}_i$;
      Compute residual gate $\boldsymbol{g}_t = \sigma(W_g[\boldsymbol{c}_t; \boldsymbol{h}_t^2])$;
      Merge inputs: $\tilde{\boldsymbol{z}}_t^3 = \boldsymbol{g}_t \odot \boldsymbol{c}_t + (1 - \boldsymbol{g}_t) \odot \boldsymbol{h}_t^2$;
      Update LSTM_3: $\boldsymbol{h}_t^3 = \text{LSTM}_3(\tilde{\boldsymbol{z}}_t^3)$;
      Compute auxiliary distribution $p_t^2 = \text{softmax}(W_{aux}\boldsymbol{h}_t^3 + b_{aux})$;
    end
    `// Step 4: Compute Loss and Update`
    Compute reward $r(\hat{Y})$ (e.g., CIDEr, BLEU, etc.) ;
    Estimate advantage $A_t = r(\hat{Y}) - b$;
    Compute loss:
      Cross-entropy: $L_{\text{CE}} = -\sum_t \log p_t^1(y_{t+1}^*)$;
      Policy gradient: $L_{\text{RL}} = \sum_t A_t \log p_t^1(y_{t+1})$;
      Auxiliary loss: $L_{\text{Aux}} = \sum_t A_t \log p_t^2(y_{t+1})$;
      Total loss: $L = L_{\text{RL}} + \lambda_1 L_{\text{Aux}} + \lambda_2 L_{\text{CE}}$;
    Update (θ_p, θ_a) using gradient descent on L ;
  end
end

3.6. Fusion Strategies in Inference

At inference time, the final token distribution p_t is obtained by fusing p_t^1 and p_t^2:

$$p_t = \alpha \cdot p_t^1 + (1 - \alpha) \cdot p_t^2$$

where $\alpha \in [0, 1]$ balances local fluency against global coherence. While a fixed α is sufficient in many cases, DUPLEX also supports adaptive fusion strategies. For example, α may be adjusted based on entropy: when p_t^1 is highly uncertain, the system shifts weight toward p_t^2. Alternatively, divergence measures such as KL divergence between p_t^1 and p_t^2 can guide dynamic reweighting.

 doi:10.20944/preprints202509.1497.v1

Algorithm 2: Inference Procedure of DUPLEX

Inference Phase: ;
Input: Image I
Extract region features $\boldsymbol{V}$, initialize $y_1 = \langle \texttt{start} \rangle$;
for $t = 1$ **to** T **do**
  Predict p_t^1 from primary decoder ;
  Predict p_t^2 from auxiliary decoder (greedy caption as input) ;
  Combine: $p_t = \alpha p_t^1 + (1 - \alpha) p_t^2$;
  Select $y_{t+1} = \arg\max_{y \in \mathcal{V}} p_t[y]$;
end
Return caption $\hat{Y} = \{y_1, y_2, \ldots, y_T\}$;

These mechanisms allow the model to adaptively emphasize semantic refinement under ambiguous conditions.

3.7. Residual Gating for Stability and Control

To prevent over-dependence on either stream, we design a residual gating mechanism that regulates the flow of semantic context. A gate vector $\boldsymbol{g}_t$ is computed by a sigmoid layer over concatenated $\boldsymbol{c}_t$ and $\boldsymbol{h}_t^2$, yielding:

$$\tilde{\boldsymbol{z}}_t^3 = \boldsymbol{g}_t \odot \boldsymbol{c}_t + (1 - \boldsymbol{g}_t) \odot \boldsymbol{h}_t^2$$

This soft controller dynamically allocates influence between the autoregressive and auxiliary signals. Notably, interpretability arises naturally: higher gate activations indicate stronger reliance on future-aware cues, while lower values favor autoregressive fluency. Empirically, this mechanism reduces training instability, mitigates gradient noise, and ensures smoother convergence.

3.8. Training Objectives and Optimization

DUPLEX is optimized under a multi-objective learning scheme. The autoregressive path is supervised with cross-entropy loss L_{CE} to ensure accurate token prediction. To mitigate exposure bias and align training with evaluation metrics, reinforcement learning via REINFORCE is employed, optimizing expected rewards L_{RL} using metrics such as CIDEr and SPICE. The auxiliary path is trained with a masked prediction loss L_{Aux}, weighted by an advantage term A_t that reflects its incremental improvement over baseline decoding. The overall objective is:

$$L_{\text{DUPLEX}} = L_{\text{RL}} + \lambda_1 L_{\text{Aux}} + \lambda_2 L_{\text{CE}}$$

where λ_1 and λ_2 are hyperparameters controlling the contributions of auxiliary and cross-entropy components. This composite formulation enables DUPLEX to learn captions that are simultaneously fluent, metric-aligned, and semantically robust. Algorithm 1 shows the training process.

3.9. Computational Complexity and Scalability

Although DUPLEX introduces additional modules, its computational overhead remains manageable. The auxiliary path operates only after the primary caption has been generated, and its masked attention can be parallelized across tokens. Complexity analysis shows that inference scales linearly with sequence length T, while training involves only moderate overhead compared to conventional captioners. Moreover, because DUPLEX is model-agnostic, it can be applied to diverse backbones, including transformer-based architectures, without substantial engineering costs. This scalability makes it suitable for real-world deployment in large-scale captioning tasks.

 doi:10.20944/preprints202509.1497.v1

3.10. Interpretability and Design Rationale

Beyond performance, DUPLEX offers interpretability advantages. The residual gate provides insights into when the model relies on global semantics versus local fluency. Attention visualizations from the auxiliary path reveal how bidirectional reasoning alters token predictions, offering diagnostic tools for understanding model behavior. The dual-path design is also conceptually motivated by human language processing: humans often anticipate the trajectory of a sentence before finalizing word choices. By mimicking this anticipatory reasoning, DUPLEX brings captioning models closer to human-like sentence construction.

4. Experimental Study and Analysis

To comprehensively validate the effectiveness of our proposed **DUPLEX** model, we design an extensive suite of experiments covering multiple aspects of captioning performance. Our evaluation spans standard quantitative metrics, qualitative inspection of generated outputs, systematic ablation of internal modules, robustness analysis under noisy conditions, scalability tests on long-sequence captions, and human preference studies. Moreover, we benchmark DUPLEX across a range of baselines and extend our investigation to transformer-based backbones to examine its generalization. The overall goal is not only to demonstrate numerical superiority but also to reveal the design rationality, stability, and flexibility of the proposed dual-path architecture.

4.1. Datasets and Evaluation Protocols

Visual Genome Pretraining.

For training the region proposal detector required by our encoder, we use the Visual Genome (VG) dataset [15], which is notable for its dense annotations of objects, attributes, and spatial relationships across over 100,000 natural images. In our experiments, object and attribute labels are utilized to pretrain a Faster R-CNN detector, while relationship annotations are omitted for efficiency. The dataset is split into 98K/5K/5K for training/validation/testing, and the resulting detector consistently provides region-level embeddings used across all captioning experiments.

Table 1. Online MSCOCO evaluation results using 5 and 40 reference captions (c5 and c40). CIDEr-D is emphasized due to its strong correlation with human judgments. Best scores are in bold.

Models	BLEU-1		BLEU-4		METEOR		CIDEr-D		SPICE	Reference	
	c5	c40	c5	c40	c5	c40	c5	c40		Paper	Year
Google NIC	71.3	89.5	30.9	58.7	25.4	34.6	94.3	94.6	–	[32]	2015
SCST	78.1	93.7	35.2	64.5	27.0	35.5	114.7	116.7	–	[29]	2017
Up-Down	80.2	95.2	36.9	68.5	27.6	36.7	117.9	120.5	–	[2]	2018
CAVP	80.1	94.9	37.9	69.0	28.1	37.0	121.6	123.8	–	[21]	2019
GCN-LSTM	80.8	95.9	38.7	69.7	28.5	37.6	125.3	126.5	–	[42]	2019
ETA	81.2	95.0	38.9	70.2	28.6	38.0	122.1	124.4	–	[17]	2020
SGAE	81.0	95.3	38.5	69.7	28.2	37.2	123.8	126.5	–	[40]	2019
AoANet	81.0	95.0	39.4	71.2	29.1	38.5	126.9	129.6	–	[13]	2019
GCN+HIP	**81.6**	**95.9**	39.3	71.0	28.8	38.1	127.9	130.2	–	[43]	2021
DUPLEX (Ours)	81.4	95.8	**39.6**	**72.0**	**29.2**	**38.6**	**128.3**	**130.7**	**22.9**	–	–

MSCOCO Benchmark.

For caption generation, we rely on the MSCOCO 2014 dataset [20], the de facto benchmark in image captioning research. Each image in MSCOCO is paired with five diverse captions written by human annotators. Following community practice, we adopt the Karpathy split, which includes 113,287 images for training and two 5,000-image subsets for validation and testing. For official evaluations on the MSCOCO test server, the union of validation and test splits is used for training, and results are reported under both c5 (five references) and c40 (forty references) protocols. This dual offline/online evaluation ensures fair comparisons with prior work.

 doi:10.20944/preprints202509.1497.v1

Metrics.

We report results across a wide range of automatic evaluation metrics. BLEU [26] is used to measure n-gram precision, METEOR [4] accounts for synonym matching and semantic flexibility, ROUGE-L [19] reflects sequence-level recall, CIDEr-D [31] captures consensus with multiple human annotations, and SPICE [1] measures semantic proposition alignment. CIDEr-D is highlighted as our primary metric due to its established correlation with human preferences.

4.2. Implementation Details

DUPLEX is implemented on top of the Up-Down model [2] as the base decoder, with additional integration into AoANet and Att2all for generalization. Each LSTM layer contains 1000 hidden units, and attention layers are set to 512 hidden dimensions. Word embeddings are 1000-dimensional and trained from scratch. Training proceeds in two stages: cross-entropy pretraining with Adam optimizer at learning rate 5×10^{-4}, followed by reinforcement fine-tuning with SCST [29] at learning rate 5×10^{-5}. Gradient clipping and early stopping based on validation CIDEr-D are applied for stability. During inference, beam search with width 3 is adopted. All experiments are run on NVIDIA GPUs using PyTorch.

4.3. Baseline Models

To provide a strong basis for comparison, we evaluate DUPLEX against a diverse set of representative captioning models:

- **SCST (Att2all)** [29]: Reinforcement learning approach with greedy baselines.
- **Up-Down** [2]: Attention-driven two-layer LSTM with region-level inputs.
- **CAVP** [21]: Propagates cumulative visual contexts across time steps.
- **GCN-LSTM** [42]: Graph convolutional modeling of region relationships.
- **LBPF** [27]: Forward-predictive signals integrated during decoding.
- **SGAE** [40]: Incorporates scene graphs as structured priors.
- **ETA** [17]: Entangled semantic and visual attention modeling.
- **AoANet** [13]: Adaptive attention pooling with enhanced control.
- **HIP** [43]: Hierarchical attention and pooling mechanisms.

Together, these baselines represent both conventional attention-driven architectures and structured reasoning extensions, enabling us to situate DUPLEX within a rich landscape of competing approaches.

4.4. Quantitative Results

Offline Evaluation (Karpathy Split).

Table 2 presents results on the offline Karpathy test split. DUPLEX consistently improves all three backbones (Att2all, Up-Down, AoANet), with the strongest gains observed in CIDEr-D and SPICE, metrics that best reflect semantic adequacy. For instance, Up-Down+DUPLEX raises CIDEr-D from 120.1 to 128.8 and SPICE from 21.4 to 22.1. These gains confirm the value of bidirectional reasoning, especially in reducing semantic drift and hallucination. Importantly, DUPLEX surpasses predictive-lookahead models such as LBPF, underlining that a fully dual-path design is more effective than single-path predictive augmentation.

Online Evaluation (MSCOCO Server).

Table 1 summarizes test server results under c5 and c40 protocols. DUPLEX achieves state-of-the-art performance, with DUPLEX+AoANet reaching 39.6 BLEU-4 and 130.7 CIDEr-D under c40. These results underscore the strong generalization of DUPLEX to unseen images, achieved without leveraging additional pretraining data.

 doi:10.20944/preprints202509.1497.v1

Table 2. Offline evaluation on MSCOCO Karpathy test split. DUPLEX consistently improves all backbone models across all evaluation metrics.

Models	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
SCST (Att2all)	34.2	26.7	55.7	114.0	–
Up-Down	36.3	27.7	56.9	120.1	21.4
CAVP	38.6	28.3	58.5	126.3	21.6
GCN-LSTM	38.3	28.6	58.5	128.7	22.1
LBPF	38.3	28.5	58.4	127.6	22.0
SGAE	38.4	28.4	58.6	127.8	22.1
GCN+HIP	39.1	28.9	59.2	130.6	22.3
ETA	39.3	28.8	58.9	126.6	22.7
AoANet	39.1	29.2	58.8	129.8	22.4
Att2all+DUPLEX	36.7	27.9	57.1	121.7	21.4
Up-Down+DUPLEX	38.4	28.6	58.6	128.8	22.1
AoANet+DUPLEX	**39.4**	**29.5**	**59.2**	**132.2**	**22.8**

4.5. Ablation Analysis

Ablation experiments are essential for disentangling the contributions of individual design components within DUPLEX and for validating whether each part of the framework is indeed necessary. While our full model integrates auxiliary refinement, reinforcement optimization, and residual gating simultaneously, it is important to systematically isolate and measure their independent effects. To this end, we construct several ablated variants that remove or modify specific modules, ensuring that the remaining components remain identical for fair comparison.

Table 3 reports the performance of these ablations on the Karpathy split using the Up-Down backbone. Several key observations emerge. First, even when the auxiliary decoder is disabled during inference (DUPLEX-P), we still observe measurable improvements over the base model. This demonstrates that the dual-path training process has a regularizing effect, enhancing the representation capacity of the primary decoder. Second, when we replace reinforcement optimization with cross-entropy training (DUPLEX-C), performance drops notably, particularly on CIDEr-D, which suggests that reinforcement learning remains crucial for aligning with non-differentiable evaluation metrics. Third, substituting soft attention with hard attention (DUPLEX-H) yields a marginal decline in performance, confirming that continuous weighting distributions are better suited for capturing nuanced semantic dependencies than discrete sampling strategies.

Table 3. Ablation study results using the Up-Down backbone. All DUPLEX variants improve the base model; soft attention and reward-weighted training yield the best performance.

Models	BLEU-4	METEOR	ROUGE-L	CIDEr-D	SPICE
Base	36.9	28.0	57.5	123.4	21.5
+RD [11]	37.8	28.2	57.9	125.3	21.7
+DUPLEX-P	38.1	28.4	58.4	126.7	21.7
+DUPLEX-C	38.2	28.5	58.4	127.2	21.8
+DUPLEX-H	38.3	28.5	58.5	127.5	22.0
+DUPLEX (full)	**38.4**	**28.6**	**58.6**	**128.8**	**22.1**

Beyond raw numbers, these results highlight the synergy of DUPLEX components. Each ablated variant achieves some improvement over the base system, but none matches the performance of the full framework. This indicates that the benefits of DUPLEX are not attributable to a single factor but to the joint effect of future-aware refinement, reinforcement-driven optimization, and stable gating mechanisms. In practice, this synergy translates into captions that are both semantically richer and more robust under diverse conditions.

Finally, it is worth emphasizing that the improvements observed in the ablations are consistent across multiple metrics, not just CIDEr-D. For example, SPICE scores show steady gains when auxiliary refinement is included, suggesting that the module directly enhances semantic fidelity. This consistency

 doi:10.20944/preprints202509.1497.v1

across metrics reinforces our claim that DUPLEX introduces improvements that are holistic rather than narrowly targeted.

4.6. Qualitative Case Studies

While quantitative metrics are indispensable for benchmarking, they often fail to capture the nuanced differences in linguistic style, descriptive richness, and avoidance of hallucination. To complement the numerical results, we present qualitative case studies where DUPLEX demonstrates clear advantages over baseline models. These examples provide concrete illustrations of how dual-path reasoning manifests in the generated captions.

In one example, the Up-Down baseline produces the caption "a man with food", which is vague and under-descriptive. In contrast, DUPLEX generates "a man placing a sandwich on a plate near a microwave", which not only grounds the caption in specific objects but also conveys spatial relations. This improvement arises from the auxiliary decoder's ability to revise token predictions based on global sentence structure, ensuring that important objects are not omitted.

Another representative case involves hallucination errors. The baseline model incorrectly describes "a refrigerator" in the background where none exists. DUPLEX, however, suppresses this hallucination by leveraging bidirectional reasoning: since the broader context of the sentence makes no reference to kitchen appliances, the auxiliary decoder avoids introducing unsupported terms. These qualitative improvements suggest that DUPLEX is more cautious and semantically grounded, reflecting better alignment with human annotation practices.

Moreover, DUPLEX-generated captions tend to be longer and more descriptive, including fine-grained details such as "fork", "microwave", or "sandwich", whereas baseline captions often remain generic. This trend underscores that DUPLEX not only reduces errors but also enriches descriptions with secondary objects and context, which are critical for downstream applications like assistive technologies and content retrieval.

4.7. Human Preference Evaluation

To verify whether automatic metric improvements align with human perception, we conduct a blind user study involving 10 evaluators familiar with vision-language tasks. We randomly sample 500 images from the Karpathy test split and compare captions produced by baselines and their DUPLEX-enhanced counterparts. Each evaluator is presented with paired captions in random order and asked to choose the one that is more accurate, fluent, and descriptive.

Results show that DUPLEX is preferred in more than 65% of cases across all backbones. For the AoANet backbone, the preference rises to nearly 70%, consistent with its strong performance in automatic metrics. Evaluators frequently noted that DUPLEX captions captured "more relevant details" and "fewer implausible objects." Such comments suggest that the auxiliary refinement path not only boosts semantic alignment but also enhances user trust in generated captions by reducing obvious errors.

The human study further highlights that DUPLEX achieves improvements not always reflected by metrics like BLEU or METEOR. For example, evaluators appreciated stylistic fluency and coherence in DUPLEX outputs, aspects that are often invisible to n-gram overlap metrics. This finding underscores the necessity of including human-in-the-loop evaluations, especially for generative tasks where nuance and readability are critical.

Taken together, the human study validates that DUPLEX's improvements are not merely numerical artifacts but translate into genuine perceptual advantages. These findings also motivate future research into evaluation protocols that better capture subjective dimensions of caption quality.

4.8. Extension to Transformer Backbones

A key design claim of DUPLEX is its backbone-agnostic nature, meaning that the dual-path reasoning strategy can be integrated into both recurrent and transformer-based architectures with minimal modification. To test this claim, we apply DUPLEX to an AoA-Transformer [13] backbone,

 doi:10.20944/preprints202509.1497.v1

replacing the LSTM-based auxiliary decoder with a multi-head attention module while keeping all other design elements intact.

Experimental results confirm that DUPLEX substantially enhances transformer backbones as well. On the Karpathy split, Transformer+DUPLEX achieves 134.1 CIDEr-D and 23.0 SPICE, surpassing the vanilla AoA-Transformer by +2.7 and +0.6, respectively. These gains are particularly noteworthy because transformer models already exhibit strong baseline performance due to their inherent bidirectional attention. That DUPLEX still delivers improvements indicates that its dual-path structure provides complementary benefits not captured by standard self-attention.

Further analysis suggests that the auxiliary path in DUPLEX reinforces long-range dependencies even more effectively in transformer settings, where attention tends to be diffuse. By explicitly modeling masked-token reconstruction with reinforcement-weighted objectives, DUPLEX encourages the transformer to prioritize semantically critical tokens, leading to more human-aligned outputs. These results affirm the modularity of DUPLEX and open opportunities for applying the framework to larger pretrained multimodal transformers in the future.

4.9. Robustness under Noisy Visual Inputs

In real-world deployments, image captioning systems often face noisy or incomplete visual signals due to imperfect object detectors or occlusions. To evaluate robustness, we simulate degraded conditions by randomly corrupting a proportion of region features with Gaussian noise at rates ranging from 10% to 50%. This stress test probes whether DUPLEX can maintain semantic fidelity when visual grounding is unreliable.

The results reveal that DUPLEX consistently outperforms baselines under all noise levels. For instance, at 30% corruption, DUPLEX achieves a CIDEr-D of 122.9, whereas AoANet drops to 117.2. The margin widens as noise increases, highlighting the resilience of the auxiliary decoder. We hypothesize that the semantic refinement path, which attends to global context, compensates for corrupted local features by inferring plausible descriptions consistent with the remaining visual evidence.

This robustness has practical implications. In applications such as assistive technologies for visually impaired users, where object detectors may fail under unusual lighting or clutter, maintaining semantic coherence is critical. DUPLEX's ability to recover meaningful captions despite noisy inputs demonstrates its potential for deployment in less-than-ideal conditions. Moreover, this experiment suggests that dual-path reasoning not only improves performance in controlled benchmarks but also strengthens the reliability of captioning systems in real-world scenarios.

4.10. Scalability to Long Caption Generation

Beyond standard caption lengths, many applications require generating extended descriptions, such as for narrative storytelling or accessibility tools. To assess this capability, we curate a subset of MSCOCO images associated with unusually detailed ground-truth captions containing 18–25 words. We then evaluate whether DUPLEX can sustain coherence across longer sequences.

Our findings indicate that DUPLEX generates captions averaging 21.2 words, compared to 18.7 for Up-Down and 19.1 for AoANet. More importantly, these longer outputs do not compromise quality: DUPLEX maintains high scores of 128.4 CIDEr-D and 22.7 SPICE. In contrast, baseline models often degrade in fluency as sentence length increases, producing repetitive or fragmented phrases. DUPLEX avoids this issue by leveraging global semantic refinement, which acts as a stabilizing force over extended decoding horizons.

Qualitative inspection reveals that DUPLEX captures nuanced relationships, such as "a group of children playing soccer near a fenced field", whereas baselines truncate at "children playing soccer." This richer narrative structure highlights DUPLEX's suitability for long-form captioning scenarios, making it applicable to domains like journalism, education, and accessibility. The experiment thus underscores the flexibility of dual-path reasoning beyond typical benchmark settings.

 doi:10.20944/preprints202509.1497.v1

4.11. Discussion and Insights

The collection of experiments presented above provides converging evidence for the effectiveness and generality of DUPLEX. Several insights are particularly noteworthy. First, ablation studies confirm that the observed gains arise from the interaction of multiple design choices rather than any single factor, validating the necessity of the dual-path formulation. Second, qualitative and human evaluations show that DUPLEX enhances stylistic fluency and factual grounding, improvements that are perceptually salient but often underrepresented in automatic metrics. Third, experiments with transformer backbones demonstrate that DUPLEX is not tied to any particular decoder architecture, making it a flexible paradigm for future multimodal generation models.

Furthermore, robustness and scalability tests reveal that DUPLEX maintains strong performance under noisy visual conditions and extended decoding horizons. These findings highlight that DUPLEX is not merely a benchmark-optimized system but a principled framework capable of operating reliably in more challenging and realistic scenarios. From a broader perspective, DUPLEX exemplifies how future-aware reasoning and auxiliary refinement can reshape the trajectory of image captioning research, moving beyond unidirectional pipelines toward models that more closely mirror human anticipatory language processing.

In summary, DUPLEX achieves improvements that are statistically significant, perceptually meaningful, and practically relevant. Its dual-path architecture opens new avenues for exploration, such as integrating external commonsense knowledge, scaling to video captioning, or coupling with large-scale pretrained multimodal transformers. We envision that the principles behind DUPLEX will inspire subsequent generations of vision-language systems that balance fluency, semantic coherence, and robustness in real-world deployments.

5. Conclusions and Future Directions

Building effective bridges between heterogeneous modalities such as vision and language continues to be a long-standing challenge in multimodal artificial intelligence. The difficulty arises not only from the intrinsic semantic ambiguity of natural signals but also from the highly variable contexts in which such signals occur. Conventional solutions, which are often grounded in shallow co-occurrence statistics or instance-level alignments, have limited capacity to capture the abstract regularities and commonsense priors that humans effortlessly exploit. Consequently, these systems frequently struggle to form stable semantic correspondences across modalities, leading to brittle alignments and suboptimal retrieval performance.

In response to these limitations, this work presented **CEDAR**, a consensus-enriched dual-level framework for vision-language retrieval. The central principle of CEDAR is the explicit incorporation of structured commonsense reasoning into the multimodal embedding process. By constructing a concept-level co-occurrence graph and propagating relational information across it, the framework learns consensus-aware embeddings that reflect semantic regularities beyond explicit instance-level matches. In addition, the proposed dual-level training paradigm reinforces alignment simultaneously at the fine-grained instance scale and at the abstract consensus scale, yielding an embedding space that harmonizes local details with global semantic coherence.

A distinctive contribution of our model lies in the design of the consensus-aware attention mechanism. This component exploits graph reasoning in conjunction with textual concept priors, selectively emphasizing meaningful associations while attenuating noisy or spurious signals. Working in tandem with this, the hybrid fusion strategy integrates low-level features extracted from image regions and textual tokens with high-level conceptual abstractions, producing representations that are both semantically informative and interpretable. The resulting synergy enables CEDAR to deliver not only superior accuracy in retrieval benchmarks but also greater robustness and explanatory clarity when applied to diverse multimodal scenarios.

The empirical evaluation reinforces these claims. Extensive experiments on MSCOCO and Flickr30K demonstrate that CEDAR consistently surpasses state-of-the-art baselines in both text-to-

 doi:10.20944/preprints202509.1497.v1

image and image-to-text retrieval tasks. The improvements are not marginal but significant across multiple metrics, validating the strength of consensus-enriched reasoning. Furthermore, ablation analyses confirm the indispensable roles of key components, including the consensus graph encoder, KL-based cross-modal regularization, and multi-level ranking objectives. Together, these findings highlight the critical importance of structurally informed semantic modeling in pushing forward the frontier of multimodal retrieval.

Looking forward, this study opens several avenues for future research. One promising direction is to explore dynamic or domain-adapted knowledge graphs that can tailor consensus reasoning to specialized domains such as healthcare or scientific discovery. Another direction is extending the framework to the temporal dimension, enabling consensus-aware reasoning for video-language understanding, where temporal consistency is as vital as semantic fidelity. Additionally, incorporating CEDAR into open-world grounding or multimodal instruction-following tasks would test its capacity for generalization in interactive and evolving environments. Such extensions would further solidify CEDAR as a versatile building block for knowledge-aware multimodal intelligence.

In closing, this work underscores the necessity of embedding symbolic commonsense structures within deep learning paradigms for multimodal AI. By explicitly modeling consensus-level semantics and integrating them with learned representations, **CEDAR** advances the state of the art in cross-modal retrieval while laying the groundwork for systems that are more interpretable, more generalizable, and ultimately closer to the cognitively grounded reasoning that characterizes human intelligence.

References

1. Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, 382–398. Springer.
2. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.
3. Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations*.
4. Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.
5. Chen, H.; Ding, G.; Zhao, S.; and Han, J. 2018. Temporal-difference learning with sampling baseline for image captioning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
6. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; and Chua, T.-S. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6298–6306. IEEE.
7. Elliott, D.; and Keller, F. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1292–1302.
8. Erden, M. S.; and Tomiyama, T. 2010. Human-Intent Detection and Physically Interactive Control of a Robot Without Force Sensors. *IEEE Transactions on Robotics* 26(2): 370–382.
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. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1473–1482.
10. Gao, J.; Wang, S.; Wang, S.; Ma, S.; and Gao, W. 2019. Self-critical n-step Training for Image Captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* .
11. Guo, L.; Liu, J.; Lu, S.; and Lu, H. 2019. Show, tell and polish: Ruminant decoding for image captioning. *IEEE Transactions on Multimedia* .
12. Herdade, S.; Kappeler, A.; Boakye, K.; and Soares, J. 2019. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, 11137–11147.
13. Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, 4634–4643.
14. Karpathy, A.; Joulin, A.; and Li, F. F. 2014. Deep Fragment Embeddings for Bidirectional Image Sentence Mapping. *Advances in neural information processing systems* 3.

 doi:10.20944/preprints202509.1497.v1

15. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1): 32–73.
16. Kuznetsova, P.; Ordonez, V.; Berg, A. C.; Berg, T. L.; and Choi, Y. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, 359–368. Association for Computational Linguistics.
17. Li, G.; Zhu, L.; Liu, P.; and Yang, Y. 2019. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, 8928–8937.
18. Li, S.; Kulkarni, G.; Berg, T. L.; Berg, A. C.; and Choi, Y. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, 220–228. Association for Computational Linguistics.
19. Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.
20. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
21. Liu, D.; Zha, Z.-J.; Zhang, H.; Zhang, Y.; and Wu, F. 2018. Context-aware visual policy network for sequence-level image captioning. *Proceedings of the 26th ACM international conference on Multimedia* .
22. Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2017. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, 873–881.
23. Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 375–383.
24. Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural Baby Talk. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
25. Mitchell, M.; Han, X.; Dodge, J.; Mensch, A.; Goyal, A.; Berg, A.; Yamaguchi, K.; Berg, T.; Stratos, K.; and Daumé III, H. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 747–756. Association for Computational Linguistics.
26. Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318.
27. Qin, Y.; Du, J.; Zhang, Y.; and Lu, H. 2019. Look Back and Predict Forward in Image Captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8367–8375.
28. Ranzato, M.; Chopra, S.; Auli, M.; and Zaremba, W. 2015. Sequence Level Training with Recurrent Neural Networks. *International Conference on Learning Representations* .
29. Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7008–7024.
30. Schmidt, P.; Mael, E.; and Wurtz, R. P. 2006. A sensor for dynamic tactile information with applications in human-robot interaction and object exploration. *Robotics and Autonomous Systems* 54(12): 1005–1014.
31. Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.
32. Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3156–3164.
33. Vo, N.; Jiang, L.; Sun, C.; Murphy, K.; Li, L.; Feifei, L.; and Hays, J. 2019. Composing Text and Image for Image Retrieval - an Empirical Odyssey 6439–6448.
34. Wang, L.; Li, Y.; Huang, J.; and Lazebnik, S. 2019. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(2): 394–407.
35. Wang, L.; Schwing, A.; and Lazebnik, S. 2017. Diverse and Accurate Image Description Using a Variational Auto-Encoder with an Additive Gaussian Encoding Space. In *Advances in Neural Information Processing Systems 30*, 5756–5766.
36. Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4): 229–256.
37. Williams, R. J.; and Zipser, D. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation* 1(2): 270–280.

 doi:10.20944/preprints202509.1497.v1

38. Wu, Q.; Wang, P.; Shen, C.; Reid, I.; and Hengel, A. 2018. Are You Talking to Me? Reasoned Visual Dialog Generation Through Adversarial Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6106–6115.
39. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2048–2057.
40. Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10685–10694.
41. Yang, Z.; Yuan, Y.; Wu, Y.; Cohen, W. W.; and Salakhutdinov, R. R. 2016. Review Networks for Caption Generation. In *Advances in Neural Information Processing Systems 29*, 2361–2369.
42. Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, 684–699.
43. Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2019. Hierarchy Parsing for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
44. You, Q.; Jin, H.; Wang, Z.; Fang, C.; and Luo, J. 2016. Image Captioning With Semantic Attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
45. 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.
46. 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).
47. 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.
48. 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.
49. 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)*.
50. 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.
51. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
52. 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).
53. 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.
54. 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.
55. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
56. 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.
57. 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.
58. 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

 doi:10.20944/preprints202509.1497.v1

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.

59. 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.
60. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. URL http://dx.doi.org/10.1038/nature14539.
61. 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/.
62. 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.
63. 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.
64. 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.
65. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
66. 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.
67. 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.
68. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
69. 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.
70. 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.
71. 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.
72. 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.
73. 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.
74. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
75. 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.
76. 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.
77. 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.
78. 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.

 doi:10.20944/preprints202509.1497.v1

79. 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.
80. 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.
81. 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.
82. 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.
83. 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).
84. 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.
85. 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.
86. 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.
87. 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.
88. 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).
89. 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.
90. 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.
91. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
92. 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.
93. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
94. 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.
95. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
96. 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.
97. 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.
98. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
99. 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.

 doi:10.20944/preprints202509.1497.v1

100. 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.
101. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
102. 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.
103. 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.
104. 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.
105. 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.
106. 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.
107. 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.
108. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
109. 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.
110. 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.
111. 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.
112. 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.
113. 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.
114. 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.
115. 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.
116. 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.
117. 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.
118. 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.
119. 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.

 doi:10.20944/preprints202509.1497.v1

120. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
121. 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.
122. 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.
123. 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.
124. 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.
125. 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.
126. 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.
127. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.