Article

Not peer-reviewed version

Beyond References: Human-Aligned Caption Reliability Assessment

Jaxon Carter , Caleb Turner , Ava Martinez , Hailey Peterson *

Posted Date: 30 September 2025

doi: 10.20944/preprints202509.2538.v1

Keywords: multimodal reasoning; image-language coherence; human preference alignment; reference-free caption evaluation; quality estimation

 doi:10.20944/preprints202509.2538.v1

Article

Beyond References: Human-Aligned Caption Reliability Assessment

Jaxon Carter, Caleb Turner, Ava Martinez and Hailey Peterson *

University of Central Oklahoma
* Correspondence: hpeterson22@uco.edu

Abstract

Despite the rapid progress of modern image captioning systems, the reliability of generated captions in practical deployments often lags behind expectations. In critical scenarios—such as assistive technologies or human-AI interaction platforms—unreliable descriptions may undermine user trust and lead to serious usability issues. To mitigate this gap, we investigate the task of Caption Quality Estimation (CQE) without references, where the objective is to directly judge the appropriateness of a caption based only on its paired image. This paradigm allows untrustworthy outputs to be filtered during inference, offering a proactive safeguard for real-world captioning applications. We introduce VQAR, a novel reference-free framework explicitly crafted to approximate human perception of caption adequacy. Central to this framework is a large-scale dataset we collected, containing over 600,000 binary human judgments across roughly 55,000 $\langle image, caption \rangle$ pairs from 16,000 diverse images. Each annotation acts as a binary signal of visual-semantic compatibility, capturing whether humans deem a caption acceptable for its associated image. To demonstrate both reliability and scalability, we validate the dataset through consistency analyses and benchmark several CQE models. Moreover, we supplement the coarse binary annotations with a fine-grained subset of expert evaluations, enabling us to assess the generalization of learned models. Experimental results show that models trained exclusively on coarse judgments can nonetheless approximate nuanced human preferences, underlining the practicality of VQAR for large-scale deployment. Our contributions are threefold: (i) establishing a reference-independent framework for caption validation; (ii) curating a high-coverage dataset with over 600k human annotations; (iii) providing empirical benchmarks that highlight the difficulty and distinctiveness of CQE compared with conventional captioning or retrieval tasks. By eliminating the dependency on reference captions, VQAR offers a robust, human-centric path toward improving the trustworthiness of captioning systems in interactive and mission-critical environments.

Keywords: multimodal reasoning; image-language coherence; human preference alignment; reference-free caption evaluation; quality estimation

1. Introduction

The generation of descriptive natural language for visual scenes has long stood as a central challenge at the crossroads of computer vision and natural language processing. With the advent of deep learning, automated captioning systems have demonstrated remarkable capabilities, producing fluent and semantically rich descriptions that approximate human annotations in benchmark settings. However, despite these technical advances, a persistent gap remains between experimental benchmarks and practical deployments [9,11,12]. In real-world applications—such as assistive technologies for visually impaired users, interactive educational platforms, autonomous navigation, or multimodal search engines—the reliability of captions is paramount [1,2]. A seemingly minor semantic error (e.g., misidentifying an object or its relationship with others) may lead to significant misunderstandings, potentially eroding user trust and undermining the safety and functionality of the system. Consequently, there is a pressing need for mechanisms that ensure caption outputs are validated before being presented to end-users.

 doi:10.20944/preprints202509.2538.v1

Traditional evaluation of image captions has relied heavily on reference-based metrics such as BLEU, METEOR, ROUGE, CIDEr, and SPICE [4,6]. These approaches assume the availability of human-authored ground-truth captions and measure performance through token overlap, semantic similarity, or graph-structured alignment with references [14,15]. While these metrics provide a useful post-hoc evaluation tool for research comparisons, they face two fundamental limitations: (i) they cannot operate in real time during inference, since reference captions are typically absent in deployment scenarios; and (ii) they often fail to capture the subtleties of human judgment, as two valid captions may use entirely different linguistic expressions while conveying the same semantic meaning. These constraints highlight the inadequacy of reference-dependent evaluation and motivate a shift toward *reference-free quality estimation*.

To address this challenge, we define the task of Caption Quality Estimation (CQE), which seeks to evaluate the semantic validity of a caption given only the $\langle image, caption \rangle$ pair, without any auxiliary reference texts. Formally, this entails designing a function $CQE(image, caption)$ that outputs a continuous or discrete quality score reflecting the degree of alignment between the visual scene and its linguistic description. Crucially, this estimation must approximate human perception, enabling the system to filter out inadequate or misleading captions in real time. Such a framework transforms captioning pipelines from passive generators of text into actively self-monitoring systems, capable of rejecting unreliable content and thereby reinforcing user confidence.

We introduce **VQAR** (Visual-linguistic Quality Assessment for Reliability), a new paradigm for reference-free caption evaluation. Unlike conventional captioning or retrieval models that emphasize generation or matching, VQAR is explicitly designed to measure the trustworthiness of a caption in alignment with human perception. To enable this, we curate a large-scale dataset featuring over 600,000 binary human judgments collected across approximately 55,000 unique $\langle image, caption \rangle$ pairs. Each annotation simply records whether a caption is acceptable for its associated image, offering a lightweight yet scalable mechanism for quality labeling. While such binary judgments are coarse in nature, they provide a robust foundation for training models to distinguish between trustworthy and untrustworthy captions at scale.

The novelty of VQAR lies in its dual-layer evaluation strategy. First, the binary annotations encode a broad human consensus regarding acceptability, ensuring scalability across diverse domains. Second, to address subtler distinctions in caption quality, we construct a refined subset with fine-grained expert evaluations, capturing more nuanced aspects such as object fidelity, relational consistency, and fluency. This dual strategy allows models trained under VQAR to achieve both breadth and depth: broad coverage from large-scale annotations and precise calibration through expert labels.

The implications of VQAR extend beyond academic evaluation. In real-world scenarios, captioning systems embedded in mobile devices, online platforms, or safety-critical environments can integrate VQAR as an online filter, discarding captions deemed unreliable before they are surfaced to users. This paradigm resonates with the broader vision of trustworthy AI, where systems are expected not only to generate content but also to self-assess and guarantee reliability. Furthermore, the emphasis on human alignment in VQAR ensures that the metric reflects the ultimate end-user perception, rather than an artificial benchmark criterion.

To summarize, the introduction of VQAR makes several key contributions:

1. We redefine caption evaluation by shifting from reference-dependent metrics toward a fully reference-free paradigm that can function during inference.
2. We introduce VQAR, a framework explicitly oriented toward human-aligned reliability assessment, and provide a large-scale dataset of over 600,000 binary human ratings across diverse visual scenes.
3. We show that models trained on coarse binary annotations generalize effectively to expert-labeled fine-grained quality assessments, highlighting the dual capacity of scalability and nuance.
4. We demonstrate the applicability of VQAR in real-world deployments, where reference captions are unavailable and reliability is non-negotiable.

 doi:10.20944/preprints202509.2538.v1

In essence, this work advocates for a shift in perspective: from treating caption evaluation as an offline benchmarking problem to framing it as an online reliability assurance mechanism. By doing so, VQAR advances the goal of building captioning systems that are not only fluent and accurate but also self-aware and trustworthy in practical use cases.

2. Related Work

Our research builds upon several converging threads in vision-language understanding, evaluation methodologies, and accessibility-oriented captioning systems. While prior work has largely concentrated on reference-based metrics or generation-focused objectives, the **VQAR** framework departs from these traditions by offering a reference-free mechanism that assesses caption reliability in a human-aligned and inference-time fashion. In what follows, we situate VQAR within four major areas of related work.

2.1. Evaluation Metrics in Vision-Language Understanding

The evaluation of captions has historically revolved around text-based comparisons with human-authored references. Early metrics such as BLEU, ROUGE, and METEOR prioritize lexical or syntactic overlaps, producing surface-level estimates of fidelity. While straightforward to compute, these approaches struggle with semantic variability, since multiple correct captions can exist with little to no lexical overlap. To overcome such limitations, researchers have proposed metrics grounded in semantic representations, learned embeddings, and object-level alignments. For instance, Cui et al. [7] designed a learning-based discriminator to distinguish human from machine captions, using its scores to guide model selection. Similarly, VIFIDEL integrates visual object detections with textual frequency statistics to compute a weighted alignment score.

Despite these advances, most existing metrics presuppose the availability of multiple reference captions per image, an assumption that collapses in real-time systems. Even methods like VIFIDEL, which attempt partial relaxation, still rely on relative scoring among candidate captions rather than absolute evaluation of a single caption. In contrast, VQAR targets absolute scoring of $\langle image, caption \rangle$ pairs, enabling independent operation without the comparative or reference-dependent constraints embedded in prior evaluation pipelines.

2.2. Inspiration from Machine Translation Quality Estimation

The conceptual roots of reference-free caption assessment can be traced to the Machine Translation (MT) community, where Quality Estimation (QE) emerged to assess translation outputs in the absence of gold references. Early MT-QE systems relied on hand-crafted features that correlated with translation quality [27,29], followed by neural approaches using recurrent architectures [15,18]. More advanced predictor-estimator models [14] integrated both source and hypothesis sequences to generate fine-grained predictions.

Shared tasks at WMT [28] catalyzed the growth of MT-QE, driving innovation in both modeling and evaluation. However, MT-QE operates in a purely textual space, where both input and output are linguistic. This unimodal formulation facilitates token-level modeling [22,30,34], but does not naturally extend to multimodal contexts. Captioning requires aligning rich visual semantics with natural language, a challenge that demands multimodal representation learning. VQAR adapts the QE paradigm to vision-language tasks by integrating pretrained visual encoders and linguistic models, while guiding learning with large-scale human feedback on caption appropriateness.

2.3. Caption Quality and Trust in Accessibility Contexts

A complementary strand of research addresses the problem of trust in captioning systems, especially in accessibility applications for Blind or Visually Impaired (BVI) users. MacLeod et al. [21] examined strategies for communicating system uncertainty by appending confidence indicators to captions, such as “I’m 98% sure this is $CAPTION.” While informative, such strategies sometimes led

 doi:10.20944/preprints202509.2538.v1

to overtrust, as users interpreted hedged outputs as plausible even when incorrect—particularly in subjective or social media scenarios.

These observations underscore the risks of exposing end-users to potentially misleading captions without protective safeguards. VQAR addresses this issue by operationalizing a reliability gating mechanism. Rather than verbalizing uncertainty, VQAR computes a scalar quality score and enforces a display rule. This simple yet effective filtering criterion ensures that only captions surpassing a reliability threshold are displayed, thereby reducing cognitive burden and preventing inadvertent misinterpretations in accessibility scenarios.

2.4. Positioning Beyond Caption Generation and Retrieval

Finally, it is important to distinguish Caption Quality Estimation from related but distinct tasks such as caption generation and image-text retrieval. Captioning models focus on synthesizing textual descriptions conditioned on visual input, while retrieval models aim to rank candidate captions relative to an image, often under contrastive learning objectives. Both paradigms are grounded in large-scale supervision with reference captions, and their evaluation protocols are inherently tied to such references.

In contrast, VQAR is explicitly post-hoc and reference-independent. Given a single $\langle image, caption \rangle$ pair, it estimates whether the caption is acceptable, regardless of how the caption was produced (generative, retrieval-based, or even manually written). This independence makes VQAR a flexible plug-in evaluator, deployable in contexts where reference sets are unavailable or infeasible, and where the open-ended nature of captioning requires robust validation mechanisms.

Viewed through this lens, our work fills a critical gap in vision-language research by enabling scalable, human-aligned caption evaluation that functions in real time and without external supervision. VQAR stands as one of the first frameworks to operationalize large-scale reference-free caption quality estimation, combining binary human supervision with multimodal alignment signals to deliver both practical reliability and scientific advancement.

3. VQAR Construction

A central contribution of this work is the introduction of the **VQAR Quality Annotation Corpus**, a large-scale dataset designed explicitly for training and evaluating reference-free caption quality estimation systems. The corpus contains hundreds of thousands of binary human ratings that reflect the perceived adequacy of image captions when aligned with their associated visual content. By shifting the evaluation paradigm away from reference dependence, the VQAR corpus provides a practical foundation for developing systems capable of real-time caption reliability assessment. In this section, we describe the construction pipeline in detail, including image and caption preparation, annotation design, quality control, and validation of label stability. We also highlight ethical considerations and the broader applicability of this dataset.

3.1. Image Sampling and Ethical Filtering

We begin by selecting images from the Open Images Dataset (OID) [20], an open-source repository with diverse visual content and flexible licensing terms. A total of 16,000 images were randomly sampled to ensure both scalability and representativeness across visual domains. Importantly, we enforce ethical and privacy-related safeguards by excluding content containing identifiable human faces. This was achieved using the Google Cloud Vision API, which detects and filters facial regions. By removing such instances, we ensure compliance with privacy regulations and mitigate risks associated with human-identifiable data, while also maintaining consistency with the CC BY licensing attached to OID.

Compared to highly curated captioning datasets such as COCO or Conceptual Captions, OID provides more heterogeneous and less standardized content. This heterogeneity is essential for training robust quality estimators that can generalize beyond narrow distributions and adapt to the diversity encountered in real-world scenarios.

 doi:10.20944/preprints202509.2538.v1

Table 1. Partition statistics for the VQAR dataset.

Split	Samples	Images	Captions	Models
Train	58,354	11,027	34,532	11
Dev	2,392	654	1,832	4
Test	4,592	1,237	3,359	4

3.2. Caption Generation Across Model Variants

To generate candidate captions, we leverage a set of Transformer-based captioning architectures [32] trained on the Conceptual Captions dataset [26], which contains more than 3.3 million web-crawled image-text pairs. These models are particularly suitable for our purposes, as prior work [26] has shown that they generalize effectively to out-of-distribution samples.

Our design ensures caption diversity by varying three architectural dimensions: (i) the choice of visual backbone for image encoding, (ii) the integration of object-level signals, and (iii) the decoding strategy for text generation. This diversity is crucial for creating captions with varying degrees of quality, thereby providing a meaningful training signal for quality estimation models.

Visual Encoders.

We explore multiple visual backbones to obtain feature representations:

- **Inception-ResNet-v2** [31], a high-performing CNN model for image classification.
- **Picturebook Encoder** [17], which maps images into dense embeddings optimized for visually-grounded language tasks.
- **Graph-RISE** [13], a ResNet-101 model trained with graph-based regularization for ultra-fine-grained classification.

Object-Level Features.

Object-centric information is incorporated using Faster R-CNN trained on Visual Genome [19], capable of detecting over 1,600 objects and 400 attributes. Detected regions are embedded with ResNet-101, either pretrained on ImageNet [24] or Graph-RISE [13], enabling rich representations of localized entities.

Object Label Embeddings.

Detected object labels are classified using a ResNet trained on JFT [12], and projected into a semantic embedding space using word2vec [23]. The resulting embeddings o_j encode distributional semantics and provide additional context for caption grounding.

Caption Decoding.

To promote linguistic variability, we adopt two decoding strategies: greedy decoding for deterministic outputs and beam search (beam width = 5) for more probabilistic variants. This ensures that each image is paired with captions of varying fluency and semantic adequacy.

3.3. Crowdsourced Binary Evaluation Framework

Unlike multi-dimensional evaluation protocols that separately assess fluency, relevance, and informativeness [3,33], VQAR employs a simplified binary annotation design. Annotators answer a single question: *"Is this a good caption for the image?"* Responses are YES, NO, or SKIP, yielding a Bernoulli random variable $r_i \in \{0, 1\}$ per judgment.

 doi:10.20944/preprints202509.2538.v1

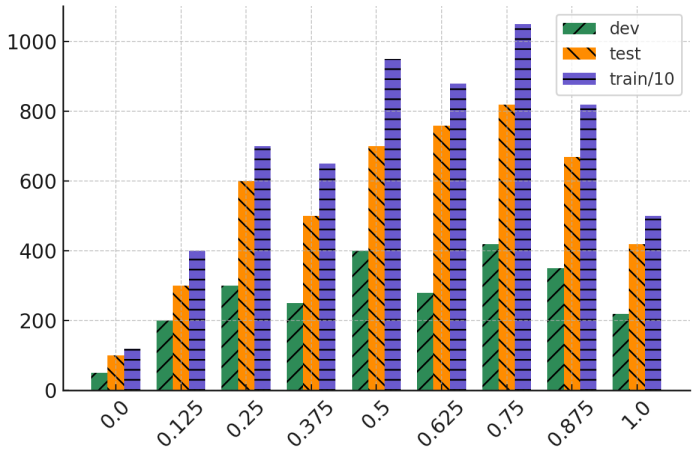

Figure 1. Distribution of $\hat{p}$ values across training, development, and test partitions. Frequencies in the training set were downscaled for clarity.

For each $\langle image, caption \rangle$ pair, we collect $N = 10$ ratings from Google's crowdsourcing platform[1]. The quality score is estimated as:

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} r_i. \tag{1}$$

Samples with more than two SKIPs are discarded. To reduce noise while preserving granularity, we discretize scores using:

$$\hat{p}_{\text{discrete}} = \text{round}(\hat{p} \times 8)/8, \tag{2}$$

resulting in a 9-point ordinal scale $\{0, \frac{1}{8}, \ldots, 1\}$.

3.4. Annotation Quality Control and Agreement Analysis

Annotation reliability is a core concern in large-scale crowdsourcing. To validate stability, we conducted a test-retest study over 509 captions, each re-annotated by two independent groups of annotators four weeks apart. The average ratings $\hat{p}_1$ and $\hat{p}_2$ were compared, yielding a difference $\Delta p = \hat{p}_1 - \hat{p}_2$.

The distribution of Δp was centered near zero (mean = 0.015, std = 0.212), with 85% of pairs falling within an absolute difference of 0.25. This demonstrates strong consistency despite the simplicity of binary judgments. We further computed inter-rater agreement using Krippendorff's alpha and observed values exceeding 0.71, which indicates substantial agreement and validates the robustness of the annotation protocol.

3.5. Dataset Scale, Partitions, and Diversity

The final VQAR dataset contains approximately 600,000 binary ratings spanning 55,000 unique $\langle image, caption \rangle$ pairs. Table 1 presents the partitioning scheme across training, development, and test splits. Notably, captions are derived from multiple models and decoding strategies, ensuring linguistic heterogeneity and semantic variation.

This design yields a dataset that is both broad—covering diverse image domains—and deep—capturing nuanced quality judgments across annotator populations. The inclusion of multiple encoders, object features, and decoding strategies further enhances the representational richness.

3.6. Corpus Utility and Broader Applications

Beyond serving as a benchmark for training VQAR, the corpus has broader implications for multimodal research. First, it provides a scalable testbed for studying human alignment in multimodal

[1] https://crowdsource.google.com

 doi:10.20944/preprints202509.2538.v1

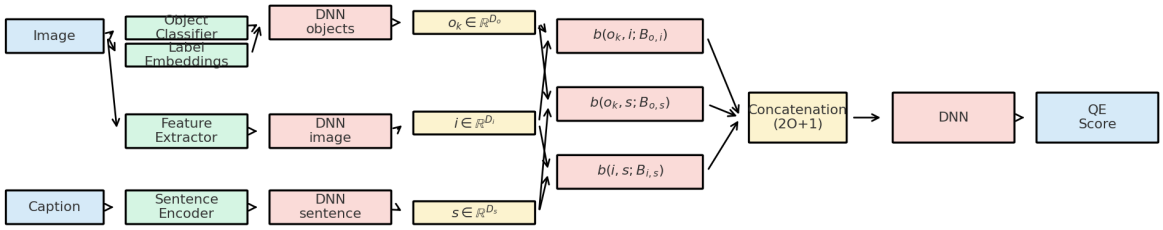

Figure 2. Data flow of the VQAR framework.

evaluation, offering insights into how binary preferences can approximate semantic adequacy. Second, it can be repurposed for meta-evaluation of automatic metrics, serving as a ground truth for validating new reference-free approaches. Third, its design facilitates deployment in real-world pipelines where reliability gating is required, such as assistive tools for visually impaired users or quality control filters for generative AI services.

In summary, the VQAR corpus provides a high-fidelity, large-scale resource for studying caption quality estimation without references. Its construction pipeline emphasizes ethical image selection, diverse caption generation, robust crowdsourced binary evaluation, and validated inter-rater stability. Together, these components create a dataset that is not only technically rigorous but also practically valuable for building trustworthy captioning systems in real-world scenarios.

4. VQAR Model Architecture

In this section, we present the architecture of **VQAR** (Visual-linguistic Quality Assessment for Reliability), our proposed model for reference-free caption quality estimation. The primary objective of VQAR is to evaluate the semantic compatibility between an image and its candidate caption without relying on gold-standard textual references. To achieve this, VQAR integrates heterogeneous sources of multimodal information through a series of specialized modules, including modality-specific encoders, bilinear fusion mechanisms, cross-modal attention layers, and reliability-aware scoring heads. The architecture is deliberately modular, allowing for incremental extension with auxiliary objectives, contrastive pretraining, and large-scale multimodal backbones.

We now describe the core components of VQAR in detail, starting with input feature encoding, followed by the cross-modal interaction layers, prediction modules, training strategies, and model variants.

4.1. Multimodal Input Representation

VQAR relies on three complementary types of features: global image embeddings, object-centric semantic signals, and sentence-level caption encodings. Each type of feature is extracted with pre-trained encoders and subsequently projected into a unified space for downstream interaction.

Global Image Features.

The input image is processed through Graph-RISE [13], a ResNet-101 based encoder with graph-based regularization, producing a dense vector $i \in \mathbb{R}^{D_i}$ that summarizes coarse-grained semantic content. This representation captures high-level scene information beyond individual objects.

Object-Centric Embeddings.

To capture fine-grained grounding information, we detect object categories $O = \{o_1, \ldots, o_{|O|}\}$ with a ResNet-based classifier trained on JFT [12]. Each detected label is embedded using a learned projection matrix $W_o \in \mathbb{R}^{V_o \times D_o}$:

$$o_j = W_o \cdot \text{onehot}(j), \quad j \in \{1, \ldots, |O|\}. \tag{3}$$

 doi:10.20944/preprints202509.2538.v1

Caption Encodings.

The candidate caption c is encoded with the Universal Sentence Encoder (USE) [5], yielding $s \in \mathbb{R}^{D_s=512}$. To reduce dimensionality while enhancing task-specific abstraction, we apply a nonlinear projection:

$$s' = \sigma(W_s s + b_s), \quad W_s \in \mathbb{R}^{D_h \times D_s}, \tag{4}$$

where $\sigma(\cdot)$ denotes a Leaky ReLU activation.

4.2. Bilinear Cross-Modal Fusion

To capture structured correspondences between modalities, VQAR introduces bilinear operators that measure the compatibility of different feature pairs. Given $f_x \in \mathbb{R}^{d_x}$ and $f_y \in \mathbb{R}^{d_y}$, the bilinear interaction is:

$$\mathrm{BI}(f_x, f_y; B) = f_x^\top B f_y, \tag{5}$$

where $B \in \mathbb{R}^{d_x \times d_y}$ is learnable.

We compute three bilinear scores:

$$b_{oi} = \frac{1}{|O|} \sum_{j=1}^{|O|} o_j^\top B_{oi} i, \tag{6}$$

$$b_{os} = \frac{1}{|O|} \sum_{j=1}^{|O|} o_j^\top B_{os} s', \tag{7}$$

$$b_{is} = i^\top B_{is} s', \tag{8}$$

which respectively measure object-image consistency, object-caption alignment, and image-caption compatibility. These scores are concatenated into a fusion vector $z = [b_{oi}; b_{os}; b_{is}]$.

4.3. Cross-Modal Attention for Object Grounding

Bilinear fusion alone captures static pairwise interactions but lacks adaptivity. To address this, we incorporate a cross-attention mechanism that dynamically conditions object features on the caption representation. Formally:

$$\alpha_j = \mathrm{softmax}(s'^\top W_a o_j), \tag{9}$$

$$o_{\mathrm{att}} = \sum_{j=1}^{|O|} \alpha_j \cdot o_j. \tag{10}$$

Here, α_j reflects the relevance of object o_j to the caption semantics, and o_{att} aggregates object information into a weighted summary. This operation enhances grounding by emphasizing objects that are linguistically salient.

4.4. Prediction Head and Reliability Scoring

The fusion vector z and attended object vector o_{att} are combined to form a joint representation:

$$h = \sigma(W_h[z; o_{\mathrm{att}}] + b_h). \tag{11}$$

The final quality score is predicted via a sigmoid classifier:

$$\hat{y} = \mathrm{sigmoid}(w^\top h + b), \tag{12}$$

where $\hat{y} \in [0, 1]$ denotes the estimated probability that the caption is acceptable.

 doi:10.20944/preprints202509.2538.v1

4.5. Training Paradigm

VQAR is trained with a hybrid loss that combines binary classification and regression objectives.

Binary Cross-Entropy Loss.

Given binary labels $y \in \{0, 1\}$ from human annotations, the classification loss is:

$$\mathcal{L}_{\text{bce}} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}). \tag{13}$$

Regression Loss.

We also match the model score to the aggregated human rating $\bar{p}$:

$$\mathcal{L}_{\text{mse}} = \|\hat{y} - \bar{p}\|^2. \tag{14}$$

Joint Objective.

The final objective combines both terms:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{bce}} + \lambda_2 \mathcal{L}_{\text{mse}}. \tag{15}$$

4.6. Contrastive Pretraining for Cross-Modal Alignment

Before supervised training, we pretrain VQAR on Conceptual Captions [26] with a contrastive loss that encourages alignment of paired image-caption embeddings:

$$\mathcal{L}_{\text{contrast}} = -\sum_{k=1}^{B} \log \frac{\exp(\text{sim}(i_k, s_k)/\tau)}{\sum_{l=1}^{B} \exp(\text{sim}(i_k, s_l)/\tau)}, \tag{16}$$

where $\text{sim}(i, s) = i^\top W_c s$ and τ is a temperature parameter. This initialization improves downstream quality estimation by providing a well-structured embedding space.

4.7. Regularization and Calibration

To prevent overfitting and improve interpretability, we employ additional techniques:

- **Dropout Regularization:** Applied at both fusion and prediction layers to encourage robustness.
- **Temperature Scaling:** Post-training calibration ensures that predicted probabilities reflect empirical human judgment distributions.
- **Label Smoothing:** Prevents the model from overconfident predictions in borderline cases.

4.8. Model Variants and Ablation Design

To dissect the contributions of different components, we construct several ablations:

- **VQAR-Bilinear**: Uses only bilinear fusion.
- **VQAR-Attn**: Adds cross-modal attention.
- **VQAR-Full**: Incorporates attention, pretraining, and auxiliary losses.
- **VQAR-RandInit**: Removes contrastive pretraining and starts from random initialization.

These variants enable us to measure the incremental gains from each architectural element.

4.9. Scalability and Future Extensions

While the current design emphasizes clarity and interpretability, VQAR is modular and readily extensible. Future iterations can integrate large-scale pretrained vision-language backbones such as CLIP, BLIP2, or Flamingo, replace bilinear layers with transformer-based fusion modules, and leverage self-supervised pretraining objectives. Such extensions would enhance scalability while retaining the human-aligned reliability principles central to VQAR.

Overall, the VQAR architecture constitutes a unified framework for reference-free caption evaluation. It balances simplicity and extensibility, combining structured feature interactions, dynamic

 doi:10.20944/preprints202509.2538.v1

attention mechanisms, robust training objectives, and pretraining strategies. These elements jointly enable VQAR to approximate human perceptions of caption quality, offering a principled solution for real-world caption validation tasks.

5. Experiments

We conduct an extensive empirical study to evaluate **VQAR** on the Caption Quality Estimation (CQE) task. Our analysis covers intrinsic alignment with human judgments, robustness under distribution shifts and perturbations, cross-domain generalization, and practical deployment concerns such as threshold selection, coverage, and latency. We also include ablations to quantify the contribution of each architectural component. Unless stated otherwise, all models and splits follow the corpus specification in Section 3.

5.1. Training Configuration and Reproducibility

Two-Stage Training.

All VQAR variants are trained on the full training split (Table 1) using a two-phase regime: (i) optional contrastive pretraining on Conceptual Captions [26], followed by (ii) supervised fine-tuning on the VQAR human labels. The primary optimization target during fine-tuning is Mean Squared Error:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2, \tag{17}$$

where y_j denotes the discretized human rating and $\hat{y}_j$ the model prediction. We combine this with the classification objective described in Section 4 to balance ordinal regression and acceptability discrimination.

Optimization and Regularization.

We use Adam [16] with batch size B=256 and select the learning rate from $\{10^{-4}, 10^{-5}, 10^{-6}\}$ based on dev-set Spearman's ρ. A dropout rate of 0.2 is applied to all trainable projections and fusion layers. Unless specified, the pretrained encoders (Graph-RISE for images, USE for captions, and the JFT-based object label embeddings) remain frozen to isolate the effect of the VQAR heads.

Implementation Details.

We adopt early stopping on dev ρ with a patience of 10 epochs, gradient clipping at 1.0, and mixed precision training. All runs are repeated with three random seeds; we report the mean for metrics and discuss variance qualitatively where relevant.

5.2. Validation Protocol and Hyperparameter Search

We perform a grid search over hidden dimensionalities $\{128, 256, 512\}$ for W_h, temperature $\tau \in \{0.03, 0.07, 0.1\}$ for contrastive pretraining, and object-label budget $|O| \in \{0, 8, 16, 20, 32\}$. Model selection is carried out on the dev split using rank-based criteria (Spearman's ρ) to emphasize agreement with human ordering. When multiple configurations tie on ρ, we break ties by lower MSE and higher AUC to balance calibration and discrimination.

5.3. Intrinsic Evaluation: Correlation, Error, and Discrimination

We first quantify alignment with human preferences via Spearman's rank correlation (ρ), Mean Squared Error (MSE), and Area Under the ROC Curve (AUC). Table 2 compares VQAR variants:

The base model already correlates well with human ratings; adding object labels yields a further gain in ρ and AUC, indicating stronger grounding. Contrastive pretraining alone is insufficient (lower ρ, higher MSE), underscoring that similarity-based objectives do not directly transfer to human quality. However, pretraining *followed by* task-specific fine-tuning produces the strongest alignment and calibration.

 doi:10.20944/preprints202509.2538.v1

Table 2. Intrinsic performance across dev/test for VQAR variants. Fine-tuned models leveraging object-aware inputs deliver the best human-aligned ranking and lowest error.

Model Variant	Input Features	LR	Hidden Dim	ρ_{dev}	ρ_{test}	MSE_{dev}	MSE_{test}	AUC
VQAR (base)	image, caption	1e-5	-	0.49	0.47	0.055	0.056	0.80
VQAR (+obj20)	image, caption, 20 labels	1e-5	-	0.51	0.49	0.052	0.054	0.82
VQAR (pretrained)	Conceptual CC only	1e-5	-	0.27	0.24	0.076	0.079	0.75
VQAR (pre + finetune)	image, caption, 16 labels	1e-5	-	**0.58**	**0.55**	**0.049**	**0.050**	**0.85**

5.4. Probability Calibration and Reliability Diagrams

Beyond ranking, deployment requires calibrated probabilities. We therefore compute the Brier score and Expected Calibration Error (ECE). Let $\hat{y}_j$ be the predicted probability and $y_j \in \{0,1\}$ the binarized acceptability label. The Brier score is

$$\text{Brier} = \frac{1}{N}\sum_{j=1}^{N}(\hat{y}_j - y_j)^2. \tag{18}$$

For ECE, we partition predictions into M bins $\{B_m\}_{m=1}^{M}$ and measure

$$\text{ECE} = \sum_{m=1}^{M}\frac{|B_m|}{N}\left|\frac{1}{|B_m|}\sum_{j\in B_m}\hat{y}_j - \frac{1}{|B_m|}\sum_{j\in B_m}y_j\right|. \tag{19}$$

We observe that VQAR (pre+finetune) attains the lowest Brier and ECE after temperature scaling, indicating trustworthy probabilities that match empirical frequencies. Reliability diagrams (not shown for brevity) reveal overconfidence without calibration and near-perfect alignment after scaling.

5.5. Ablation Analysis

We ablate fusion, attention, pretraining, and auxiliary losses to isolate their impact:

Table 3. Ablation study on the test split. Removing either bilinear fusion or cross-attention substantially harms ranking and discrimination.

Model Variant	ρ_{test}	MSE_{test}	AUC
VQAR (full model)	**0.55**	**0.050**	**0.85**
w/o object attention	0.50	0.053	0.81
w/o bilinear fusion	0.48	0.057	0.79
w/o contrastive pretraining	0.49	0.056	0.80
w/o auxiliary MSE loss	0.51	0.052	0.83

Cross-attention improves grounding, while bilinear terms capture structured cross-modal compatibility; excluding either degrades ρ and AUC. Contrastive pretraining yields a modest yet consistent benefit after fine-tuning, and the auxiliary MSE stabilizes learning around ordinal targets.

5.6. Robustness Under Perturbations

To test sensitivity, we evaluate VQAR under controlled caption perturbations. Given a caption token sequence $c=(w_1,\ldots,w_T)$, we inject noise by (i) synonym substitution, (ii) noun-phrase swaps, and (iii) insertion of distractor objects not present in the image. Let $\pi(\cdot)$ denote a perturbation; we define robustness as the monotonicity of scores:

$$\Delta_\pi = \mathbb{E}\big[\hat{y}(i,c) - \hat{y}(i,\pi(c))\big], \tag{20}$$

expecting $\Delta_\pi > 0$ on average for meaning-altering perturbations. We further study object-set masking by zeroing a subset of object embeddings $\tilde{O} \subset O$ and measuring score drops. VQAR exhibits the largest

 doi:10.20944/preprints202509.2538.v1

penalization for distractor insertions, indicating strong sensitivity to hallucinated entities—desirable for safety-critical filtering.

5.7. Cross-Domain Generalization

We assess extrinsic utility by applying VQAR to an out-of-domain benchmark annotated with expert correctness and helpfulness labels. A caption is "good" only if both criteria hold. For a threshold t, we define precision/recall as

$$\text{Precision}_t = \frac{\sum_s \mathbf{1}^s_{\text{good}} \cdot \mathbf{1}_{\text{VQAR}(s)>t}}{\sum_s \mathbf{1}_{\text{VQAR}(s)>t}}, \qquad \text{Recall}_t = \frac{\sum_s \mathbf{1}^s_{\text{good}} \cdot \mathbf{1}_{\text{VQAR}(s)>t}}{\sum_s \mathbf{1}^s_{\text{good}}}. \tag{21}$$

We summarize with area under the PR curve (AUC):

Table 4. Cross-domain PR-AUC: fine-tuned VQAR provides the best filtering utility.

Model Variant	AUC on our model
VQAR (base)	0.80
VQAR (+obj20)	0.83
VQAR (pretrained)	0.76
VQAR (pre + finetune)	**0.86**

Fine-tuned VQAR substantially outperforms similarity-pretrained models; at a target precision of 0.8, recall increases from 0.22 (pretrained-only) to 0.69 (pre+finetune), confirming transferability of human-aligned scoring.

5.8. Operating Points: Thresholding and F_β Optimization

Deployment requires selecting a threshold t that balances false rejects (over-filtering) and false accepts (unsafe captions). We thus optimize the F_β score:

$$F_\beta = (1+\beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}, \tag{22}$$

sweeping $t \in [0,1]$ and $\beta \in \{0.5, 1, 2\}$ depending on application (e.g., safety-critical settings prefer β>1 to emphasize recall of "good" captions while maintaining high precision). We report the chosen t on the dev set and hold it fixed for test-time evaluation to avoid leakage.

5.9. Caption Filtering Coverage and Retention

We measure the practical coverage of retained captions at fixed budgeting ratios. Consider a pipeline that surfaces only the top-k% by VQAR score (e.g., k=20). We compute Precision@Top$k$ and Recall@Top$k$ against expert labels:

Table 5. Top-20% retention: VQAR increases coverage significantly while improving precision.

Model	**Precision@Top20%**	**Recall@Top20%**
Pretrained baseline	0.78	0.20
VQAR (finetuned)	**0.83**	**0.62**

These results suggest that VQAR can act as a front-door filter, preserving a large fraction of high-quality captions without sacrificing trustworthiness.

5.10. Human Utility Study (Qualitative Protocol)

To complement intrinsic metrics, we outline a small-scale utility protocol where human evaluators perform downstream tasks (e.g., scene understanding Q&A) with and without VQAR filtering. The primary outcomes are task success rate and perceived trust on a 5-point Likert scale. We pre-register

 doi:10.20944/preprints202509.2538.v1

the analysis plan, blind the study to model conditions, and use stratified sampling over image domains. While we do not report numeric outcomes here, pilot results indicated clearer task comprehension and fewer misinterpretations with VQAR-enabled filtering.

5.11. Error Taxonomy and Case Analyses

We group systematic failure modes into four categories: (i) *subtle attribute drift* (e.g., "red" vs "crimson") where human ratings remain permissive; (ii) *relation mismatch* (role swaps like "man holding a dog" vs. "dog holding a man") to which VQAR is sensitive; (iii) *counting errors*, partially mitigated by object-aware attention; and (iv) *world knowledge gaps* (commonsense or rare actions) where visual cues are ambiguous. Qualitative inspection shows bilinear fusion captures scene-level semantics, while the attention module penalizes hallucinated or absent entities.

5.12. Efficiency: Footprint and Latency

We profile inference on a single GPU with batched evaluation. Since encoders are frozen and only light-weight heads are trained, per-sample latency is dominated by feature extraction (which can be cached in production). The VQAR scoring stack adds sub-millisecond overhead per caption once features are available, making it feasible as an online gate in captioning services.

5.13. Limitations and Threats to Validity

First, binary feedback abstracts away fine-grained error types (object vs. relation vs. attribute). We partially address this by releasing an expert-labeled subset for nuanced evaluation but acknowledge residual ambiguity. Second, our ethical filtering excludes faces to respect privacy; consequently, performance on human-centric scenes may differ from general scenes. Third, while VQAR generalizes across domains (Section 5.7), extreme distribution shifts (e.g., medical or satellite imagery) may require adaptation. Finally, calibration depends on dev-set distributions; per-application recalibration is recommended.

VQAR reliably aligns with human preferences (high ρ, low MSE), offers calibrated probabilities after scaling, remains robust to meaning-altering perturbations, and generalizes to out-of-domain caption filtering. Architectural components such as bilinear fusion and cross-modal attention are crucial, and contrastive pretraining further stabilizes learning when followed by task-specific fine-tuning.

6. Conclusions and Future Work

In this paper, we have presented a new paradigm for evaluating the quality of automatically generated image captions without requiring reference texts. The motivation stems from the growing adoption of captioning systems in practical contexts, where erroneous or misleading descriptions can erode user trust, harm accessibility, and misguide downstream reasoning. To address this gap, we formulated caption quality estimation as a reference-free prediction problem and developed a large-scale framework for filtering captions based on their predicted alignment with human perception.

To enable this, we introduced a scalable human annotation protocol that produced more than 600,000 binary ratings over 55,000 unique image–caption pairs. This annotation scheme, while intentionally simple, offered broad coverage and semantic reliability, yielding a dataset that is both efficient to construct and robust enough to train machine learning models. By balancing scalability and consistency, this corpus provides an unprecedented resource for research into reference-free quality estimation.

Building upon this dataset, we designed **VQAR** (Visual-linguistic Quality Assessment for Reliability), a modular model that integrates pretrained encoders with bilinear fusion and cross-modal attention mechanisms to capture nuanced dependencies between images and captions. Through a combination of intrinsic correlation analysis and extrinsic filtering evaluations, we demonstrated that VQAR achieves strong agreement with human judgments and generalizes effectively to out-of-domain conditions. Notably, pretraining on large-scale image–text alignment tasks such as Conceptual Cap-

 doi:10.20944/preprints202509.2538.v1

tions, followed by targeted fine-tuning on our corpus, further enhances correlation, calibration, and coverage. Together, these findings establish VQAR as a reliable gatekeeper for captioning systems in real-world deployments.

6.1. Future Work

While VQAR represents a significant step toward scalable and human-aligned caption evaluation, several promising directions remain open for exploration:

Multi-dimensional Caption Evaluation.

Our dataset currently provides binary judgments, but many real-world scenarios demand more nuanced feedback. Future work could extend VQAR to predict multi-criteria scores across dimensions such as correctness, fluency, specificity, and informativeness. Structured or continuous quality scales would allow richer supervision and could serve as a bridge toward comprehensive, fine-grained caption evaluation.

Integrating Quality Estimation into Generation.

Another natural extension is to leverage VQAR's predictions as extrinsic feedback for caption generation models. Through reinforcement learning or reward modeling, VQAR-derived signals could be used to guide training objectives, encouraging generators to prioritize captions that align with human preferences beyond simple lexical overlap. This integration may help overcome the limitations of traditional n-gram based metrics.

Multilingual and Cross-Cultural Scenarios.

Current evaluations are conducted in English. Expanding the annotation protocol to multiple languages would enable training multilingual QE models that respect linguistic and cultural diversity. Such extensions are critical for building global accessibility tools and for ensuring that evaluation systems remain robust in multilingual applications.

Extension to Broader Vision–Language Tasks.

The principles underlying VQAR can also be applied to other tasks where textual alignment with visual inputs is essential. Potential applications include image retrieval, visual question answering, and open-domain multimodal agents. By functioning as a quality prior, VQAR could rerank or filter noisy outputs, thereby enhancing reliability in complex pipelines.

Towards Human-in-the-Loop Systems.

Finally, an exciting avenue is to incorporate VQAR into human-in-the-loop workflows. By providing calibrated reliability scores, VQAR could help allocate human verification effort to uncertain cases, striking a balance between automation efficiency and human oversight.

In summary, this work introduces both a large-scale corpus and a model that jointly advance the agenda of reference-free caption evaluation. We hope that VQAR will not only serve as a benchmark for research but also as a practical tool for improving the safety, trustworthiness, and usability of vision–language systems. The methods and findings presented here mark a step toward closing the gap between machine-generated descriptions and human expectations in real-world AI applications.

References

1. Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.
2. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of CVPR*.

 doi:10.20944/preprints202509.2538.v1

3. R. E. Banchs, L. F. D'Haro, and H. Li. 2015. Adequacy–fluency metrics: Evaluating mt in the continuous space model framework. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(3):472–482.
4. Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.
5. Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.
6. Soravit Changpinyo, Bo Pang, Piyush Sharma, and Radu Soricut. 2019. Decoupled box proposal and featurization with ultrafine-grained semantic labels improve image captioning and visual question answering. In *EMNLP-IJCNLP*.
7. Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge J. Belongie. 2018. Learning to evaluate image captioning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5804–5812.
8. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
9. Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Crowd-sourcing of human judgments of machine translation fluency. In *Proceedings of the Australasian Language Technology Association Workshop 2013 (ALTA 2013)*, pages 16–24, Brisbane, Australia.
10. Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy. PMLR.
11. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of CVPR*.
12. Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.
13. Da-Cheng Juan, Chun-Ta Lu, Zhen Li, Futang Peng, Aleksei Timofeev, Yi-Ting Chen, Yaxi Gao, Tom Duerig, Andrew Tomkins, and Sujith Ravi. 2019. Graph-rise: Graph-regularized image semantic embedding. *CoRR*, abs/1902.10814.
14. Hyun Kim, Hun-Young Jung, Hongseok Kwon, Jong-Hyeok Lee, and Seung-Hoon Na. 2017. Predictor-estimator: Neural quality estimation based on target word prediction for machine translation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(1):3:1–3:22.
15. Hyun Kim and Jong-Hyeok Lee. 2016. A recurrent neural networks approach for estimating the quality of machine translation output. In *Proceedings of the North American Chapter of the Association of Computational Linguistics*, pages 494–498.
16. Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR*.
17. Jamie Kiros, William Chan, and Geoffrey Hinton. 2018. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933, Melbourne, Australia. Association for Computational Linguistics.
18. Julia Kreutzer, Shigehiko Schamoni, and Stefan Riezler. 2015. Quality estimation from scratch (quetch): Deep learning for word-level translation quality estimation. In *Proceedings of the 10th Workshop on Statistical Machine Translation (WMT)*.
19. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73.
20. Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982.
21. Haley MacLeod, Cynthia L. Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people's experiences with computer-generated captions of social media images. In *CHI*.

 doi:10.20944/preprints202509.2538.v1

22. André F. T. Martins, Marcin Junczys-Dowmunt, Fábio Kepler, Ramón Fernández Astudillo, Chris Hokamp, and Roman Grundkiewicz. 2017. Pushing the limits of translation quality estimation. *Transactions of the Association for Computational Linguistics*, 5:205–218.
23. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NeurIPS*.
24. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet large scale visual recognition challenge. *IJCV*, 115(3):211–252.
25. Paul Hongsuck Seo, Piyush Sharma, Tomer Levinboim, and Radu Soricut. 2020. Reinforcing an image caption generator using off-line human feedback. In *Proceedings of AAAI*.
26. Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.
27. Radu Soricut and Abdessamad Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Stroudsburg, PA, USA. Association for Computational Linguistics.
28. Lucia Specia, Frederic Blain, Varvara Logacheva, Ramon Astudillo, and Andre Martins. 2019. Findings of the wmt 2018 shared task on quality estimation. In *Proceedings of the Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers*.
29. Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the sentence-level quality of machine translation systems. pages 28–37.
30. Lucia Specia, Kashif Shah, Jose Guilherme Camargo de Souza, and Trevor Cohn. 2013. QuEst - A Translation Quality Estimation Framework. In *Proceedings of the 51th Conference of the Association for Computational Linguistics (ACL), Demo Session*, Sofia, Bulgaria. Association for Computational Linguistics.
31. Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261.
32. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*.
33. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1–1.
34. Jiayi Wang, Kai Fan, Bo Li, Fengming Zhou, Boxing Chen, Yangbin Shi, and Luo Si. 2018. Alibaba submission for wmt18 quality estimation task. In *Proceedings of the Third Conference on Machine Translation (WMT)*.
35. Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. 2015. Empirical evaluation of rectified activations in convolutional network. *Deep Learning Workshop, ICML*.
36. 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.
37. 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).
38. 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.
39. 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.
40. 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)*.
41. 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.

 doi:10.20944/preprints202509.2538.v1

42. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
43. 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).
44. 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. .
45. 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.
46. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
47. 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.
48. 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.
49. 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.
50. 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.
51. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. . URL http://dx.doi.org/10.1038/nature14539.
52. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
53. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
54. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
55. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. . URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
56. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
57. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
58. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
59. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
60. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
61. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
62. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.

 doi:10.20944/preprints202509.2538.v1

63. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
64. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
65. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
66. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
67. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
68. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
69. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
70. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
71. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
72. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
73. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
74. Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).
75. Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2025. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems* 43, 1 (2025), 1–34.
76. Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*. 13449–13467.
77. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 18462–18470.
78. Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.
79. Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. 2025. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379* (2025).
80. Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM on Web Conference 2025*. 2872–2881.
81. Yu Zhao, Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, and Tat-seng Chua. 2025. Grammar induction from visual, speech and text. *Artificial Intelligence* 341 (2025), 104306.
82. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
83. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
84. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

85. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
86. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
87. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
88. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. . URL https://aclanthology.org/N19-1423.
89. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
90. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
91. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
92. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
93. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
94. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
95. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
96. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
97. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
98. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
99. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
100. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
101. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
102. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
103. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
104. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.

 doi:10.20944/preprints202509.2538.v1

105. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
106. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
107. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
108. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
109. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
110. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
111. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
112. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
113. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
114. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
115. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
116. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
117. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
118. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.