Enhancing Caption Fidelity via Explanation-Guided Captioning with Vision-Language Fine-Tuning

Luca Müller *, Rodolfo Patel, Sofia Rossi

Posted Date: 1 August 2025

doi: 10.20944/preprints202508.0076.v1

Keywords: Image Captioning; Explainability; Layer-wise Relevance Propagation; Attention Mechanisms; Hallucination Mitigation; Vision-Language Models; Gradient-based Explanation; Fine-tuning Strategies

 doi:10.20944/preprints202508.0076.v1

Article

Enhancing Caption Fidelity via Explanation-Guided Captioning with Vision-Language Fine-Tuning

Luca Müller *, Rodolfo Patel and Sofia Rossi

University of Charleston
* Correspondence: luca.m@cofc.edu

Abstract

Image captioning models have achieved remarkable progress with the introduction of attention mechanisms and transformer-based architectures. However, understanding and diagnosing their predictions remain a challenging task, particularly in terms of attribution, interpretability, and mitigation of hallucinated outputs. In this work, we present **CAPEV**, a novel explanation-guided fine-tuning paradigm that builds upon Layer-wise Relevance Propagation (LRP) to improve caption reliability and semantic grounding. We begin by systematically adapting state-of-the-art explanation methods—including LRP, Grad-CAM, and Guided Grad-CAM—to image captioning architectures with both adaptive and multi-head attention mechanisms. Unlike conventional attention heatmaps, which offer a coarse visual explanation, these gradient-based and propagation-based methods provide dual-perspective relevance: spatial pixel-level attributions for image regions and token-wise linguistic relevance across sequential inputs. Through rigorous comparisons, we find that these methods yield a more precise and disentangled understanding of the model's decision basis. Building on these insights, we introduce CAPEV, an inference-time fine-tuning approach that leverages explanation signals to recalibrate the internal representations of the model. By identifying both supporting and opposing relevance cues for each word prediction, CAPEV dynamically adjusts context features to suppress hallucinated entities and reinforce grounded content. Notably, CAPEV operates without requiring additional external annotations or human supervision. Extensive experiments on Flickr30K and MSCOCO benchmarks demonstrate that CAPEV significantly reduces object hallucination while preserving caption fluency and overall performance on standard evaluation metrics. Our findings suggest that integrating explainability into the training loop opens a promising avenue toward transparent and trustworthy vision-language generation.

Keywords: image captioning; explainability; layer-wise relevance propagation; attention mechanisms; hallucination mitigation; vision-language models; gradient-based explanation; fine-tuning strategies

1. Introduction

Image captioning—the task of automatically generating natural language descriptions from visual data—has garnered increasing attention in the field of vision-language understanding. The standard modeling paradigm typically comprises a convolutional neural network (CNN) as an image encoder and a recurrent neural network (RNN), often an LSTM or Transformer, as a text decoder [1–3]. A key advancement in this area is the introduction of attention mechanisms, which allow the decoder to selectively focus on relevant regions of the image when generating each word [4,5,8,11,12,15].

While attention heatmaps are often used as visual explanations of the model's focus, they primarily reflect a heuristic alignment between image regions and textual tokens. However, such attention maps fall short in explicitly distinguishing the respective contributions of the image input and the sequential textual history. In particular, they do not provide clarity on whether the model is actually using the image evidence or merely relying on learned language priors—a common issue in captioning tasks.

 doi:10.20944/preprints202508.0076.v1

To address these limitations, we adapt several explanation techniques from the interpretability literature—specifically, Layer-wise Relevance Propagation (LRP) [23,24], Grad-CAM, and Guided Grad-CAM [21,22]—to the image captioning domain. These methods enable high-resolution attribution across both visual and textual modalities. In particular, LRP provides signed relevance maps, indicating which features contribute positively or negatively to the model's decisions, for both image pixels and prior generated tokens.

Through qualitative and quantitative analyses, we demonstrate that these methods offer a more faithful and granular decomposition of model decisions compared to conventional attention maps. Moreover, they uncover previously hidden model behaviors—such as over-reliance on frequent textual patterns or incorrect associations between objects and words—that are often responsible for hallucinated outputs. This is aligned with prior observations on hallucination issues in captioning systems [27,29,31].

Recognizing the power of explanation-based feedback, we design a novel fine-tuning strategy—**CAPEV**—which uses LRP-derived relevance scores to modulate the context representation during inference-time learning. Unlike traditional fine-tuning approaches that adjust model weights via backpropagation with global gradients, CAPEV focuses on local relevance signals. Specifically, features with high positive relevance are amplified, while those with negative contributions are suppressed, enabling the model to align its output more precisely with the true visual evidence. Our proposed modulation mechanism is lightweight, plug-and-play, and does not require external supervision or additional annotations—unlike prior approaches that rely on curated segmentation maps or human relevance annotations [28,29].

Importantly, CAPEV also addresses the phenomenon of "semantic shortcutting", where models predict plausible but visually unsupported tokens due to dataset biases. By emphasizing grounded evidence, our method enhances both interpretability and accuracy. Experiments on MSCOCO and Flickr30K validate the effectiveness of CAPEV in improving mean average precision (mAP) of object-centric predictions while maintaining BLEU, METEOR, ROUGE, CIDEr, and SPICE scores at competitive levels [16,17,19].

To summarize, our key contributions include:

- We adapt and extend explanation methods to generate multimodal relevance for image captioning models, revealing detailed attribution over both image and language inputs.
- We perform a rigorous quantitative analysis of these explanations in terms of grounding accuracy, interpretability, and their capacity to expose hallucinations and misaligned predictions.
- We propose CAPEV, a relevance-guided fine-tuning framework that reduces hallucinated object descriptions without compromising caption fluency or requiring extra supervision.

2. Related Work

2.1. Advances in Image Captioning Architectures

Image captioning has long been a benchmark task for evaluating cross-modal understanding between vision and language. The conventional modeling paradigm follows the encoder-decoder architecture, where a convolutional neural network (CNN) is utilized to extract dense visual features and a recurrent neural network (RNN), typically an LSTM or GRU, is employed to decode these features into coherent natural language sequences [1–3]. While this pipeline captures the basic semantics of the visual scene, it struggles with fine-grained object interactions and contextual reasoning.

To alleviate this limitation, attention mechanisms have been widely adopted to selectively focus on informative regions of the image while generating each word in the caption. Early variants such as soft and hard attention models [4] laid the groundwork for later attention-based advancements. Semantic attention [6], adaptive attention [7], bottom-up and top-down attention [8], and hierarchical attention mechanisms [10] introduced refined control over region-to-word mappings.

Recent breakthroughs in sequence modeling led to the rise of Transformer-based attention [11], where self-attention mechanisms facilitate rich cross-token interactions. These include Attention-

 doi:10.20944/preprints202508.0076.v1

on-Attention (AoA) networks [12], Entangled Transformer structures [13], and Meshed-Memory Transformers [15]. These architectures incorporate multi-head attention layers that simultaneously attend to multiple semantic aspects across the image and caption tokens.

Despite their success, attention-based models still face challenges in capturing object-level relationships and contextual attributes. To address this, graph-based scene modeling has emerged as a complementary approach. Scene graphs [28,37], attribute-based graphs [38], and hierarchical relational modeling frameworks [39,40] have been proposed to enrich image representations with structured semantics. Visual Relation Graphs (VRG) [41] capture inter-object relationships and spatial dependencies, leading to more context-aware caption generation.

Beyond static object modeling, a line of work focuses on local-global representation fusion. Techniques like visual-linguistic distillation [42], noun-chunk parsing [44], and policy-based gradual representation learning [43] aim to align local details with global context. Pretrained vision-language models such as Unified VLP [45], OSCAR [48], and VIVO [47] have pushed the boundaries further by leveraging external knowledge sources, e.g., image-tag pairs, to build universal multimodal embeddings akin to BERT [46].

Parallel to the mainstream task, several challenging extensions of image captioning have been studied. Notably, novel object captioning (NOC) [50–55] targets the out-of-distribution generalization problem by enabling models to describe previously unseen objects. Other directions include controllable style transfer [56], sentiment-aware captioning [57], and human-centered captioning [49], all aimed at enhancing the expressiveness and personalization of generated language.

2.2. Mitigating Bias in Vision-Language Systems

Multimodal models are inherently susceptible to dataset-induced biases due to the co-occurrence patterns in visual and textual modalities. In particular, vision-language models often exploit superficial correlations or dominant priors during inference, leading to biased or hallucinated outputs. This issue has been widely studied in both the visual question answering (VQA) and image captioning domains.

In the VQA domain, RUBi [30] proposes to down-weight biased language features during training to force models to rely more on visual signals. Similar efforts such as [58] disentangle question types to better isolate language-induced biases and introduce visual grounding constraints.

In the context of image captioning, gender bias and object priors are among the most prominent problems. Hendricks et al. [29] introduced appearance confusion and confidence loss terms to reduce gender stereotypes, using segmentation-based annotations to supervise the de-biasing process. HINT [31] adopts human-annotated attention maps to realign model predictions with visually grounded regions, thus guiding the model to "look before it talks."

However, these methods typically require costly external annotations, such as segmentation maps or patch rankings. In contrast, the CAPEV framework introduced in our work leverages internal explanation signals generated by LRP, which inherently reflect the contribution of both visual and textual features without extra supervision. This design allows for a scalable and annotation-free approach to mitigating hallucination caused by language bias.

2.3. Interpretability and Explanation for Captioning Models

As deep neural networks become increasingly opaque, interpretability techniques offer an essential lens through which model decisions can be understood and debugged. In the domain of image captioning, the challenge is further exacerbated by the sequential and multimodal nature of the task, making attribution analysis more intricate than in single-modal tasks like image classification.

Explanation techniques can broadly be categorized into three types: (1) gradient-based methods such as saliency maps [59], Guided Backpropagation [22], Integrated Gradients [60], and Grad-CAM [21]; (2) decomposition-based methods such as Layer-wise Relevance Propagation (LRP) [23], DeepLIFT [62], and Contextual Decomposition [65]; and (3) perturbation- and sampling-based methods such as LIME [66], RISE [69], and Occlusion-based techniques [67].

 doi:10.20944/preprints202508.0076.v1

These methods have been extended to various neural architectures, including CNNs, RNNs, GNNs [72,73], and clustering models [77]. However, few works have attempted to adapt these methods for image captioning models, despite their complex multimodal dependencies. Early efforts such as [78] treated static images as videos to apply temporal explanation tools, while Grad-CAM has been applied in limited non-attention settings [21].

While attention heatmaps are often used as proxy explanations in captioning models, their interpretability has been critically questioned in NLP contexts [79,81]. Attention maps typically highlight spatial regions, but fail to convey signed relevance or disentangle modality-specific contributions. In this work, we instead adopt and extend LRP and gradient-based explanation techniques to generate pixel-level and token-level attribution signals, offering deeper insights into the behavior of captioning models.

2.4. Using Explanations to Guide Training

Recent research has explored the intersection between explainability and learning, showing that explanation signals can be used as auxiliary supervision to guide model optimization. Grad-CAM has been employed to design saliency-aware cross-entropy losses in classification [82], leading to improved robustness and visual grounding.

In image captioning, HINT [31] introduced a loss function that ranks image patches based on human annotations and explanation saliency, helping the model focus on correct visual regions during training. Similarly, Sun et al. [83] used LRP explanations in few-shot learning to identify transferable features and enhance generalization.

Building on this idea, our CAPEV framework leverages LRP-generated relevance signals not just for interpretation, but as actionable feedback to drive inference-time fine-tuning. This approach serves as a bridge between model transparency and practical performance gains, aligning the decision-making process with human-understandable evidence without introducing any external supervision cost.

3. Preliminary to Image Descriptions Captioning

3.1. Notational Framework and Pipeline Overview

We begin by formalizing the key components of typical image captioning systems, which follow an encoder-decoder architecture comprising three essential modules: a visual encoder, a language decoder, and a fusion-based word predictor.

Given an input image, a visual encoder—such as a CNN or a region-based detector like Faster R-CNN—is applied to extract spatial or region-level features denoted by $\boldsymbol{I} \in \mathbb{R}^{n_v \times d_v}$, where n_v is the number of regions or spatial locations, and d_v is the dimensionality of the feature vectors. For detectors like Faster R-CNN, n_v corresponds to the number of object proposals; for CNN feature maps, n_v corresponds to the number of flattened spatial patches.

At each decoding step t, a hidden representation $\boldsymbol{h}_t$ is computed via an LSTM conditioned on the previous word and a global visual summary. The decoder state evolves as:

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

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

Here, $\boldsymbol{E}_w(\cdot)$ denotes the word embedding function, and $\boldsymbol{I}_g = \frac{1}{n_v}\sum_{k=1}^{n_v} \boldsymbol{I}_{(k)}$ is a global average-pooled visual descriptor. The LSTM output state $\boldsymbol{h}_t$ is further used to generate attention-guided context features $\boldsymbol{c}_t$ via an attention module:

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

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

 doi:10.20944/preprints202508.0076.v1

The resulting score vector $\boldsymbol{p}_t$ yields a distribution over vocabulary words at time t. The attention function $\mathrm{ATT}(\cdot)$ and predictor module can be instantiated in various ways, leading to different model families.

3.2. Dynamic Visual-Linguistic Integration via Attention Modules

We investigate two widely adopted attention paradigms: *adaptive attention*, which leverages a gating sentinel to control visual-textual flow, and *multi-head attention*, a Transformer-based design that facilitates parallel attention to diverse semantic cues.

3.2.1. Adaptive Attention: Sentinel-Guided Integration

Adaptive attention introduces an auxiliary memory cell $\boldsymbol{s}_t$—termed the *sentinel vector*—that selectively captures textual memory separate from the visual stream. At each time step, the sentinel is updated as:

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

where $\boldsymbol{W}_x \in \mathbb{R}^{d_h \times d_x}$ and $\boldsymbol{W}_h \in \mathbb{R}^{d_h \times d_h}$ are learned projection matrices. The sigmoid gate modulates memory exposure. Subsequently, attention scores over both visual features and sentinel are computed:

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

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

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

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

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

Here, β_t balances visual versus textual contribution in the final attended vector. We define the adaptive attention operator compactly as:

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

3.2.2. Multi-Head Attention: Parallel Contextual Projections

In contrast, multi-head attention utilizes multiple linear projections of queries, keys, and values to allow the model to jointly attend to different aspects of the image. The formulation proceeds as:

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

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

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

The outputs from each attention head are concatenated and transformed:

$$\boldsymbol{v} = [\boldsymbol{v}^{(1)}, \cdots, \boldsymbol{v}^{(n_h)}] \tag{15}$$

$$\hat{\boldsymbol{v}} = \boldsymbol{W}_v \boldsymbol{v} + \boldsymbol{b}_v \tag{16}$$

To model visual contribution uncertainty, we apply a gating mechanism to form the final context:

$$\boldsymbol{c}_t = \sigma(\boldsymbol{W}_{mh} \boldsymbol{h}_t + \boldsymbol{b}_{mh}) \odot \hat{\boldsymbol{v}} = \mathrm{ATT}_{\mathrm{mha}}(\boldsymbol{h}_t, \boldsymbol{I}) \tag{17}$$

3.3. Unified Model Architectures for Evaluation

To provide a controlled analysis, we instantiate two representative captioning models:

 doi:10.20944/preprints202508.0076.v1

- **Ada-LSTM**: Incorporates the adaptive attention module alongside an LSTM decoder and a fully connected prediction head.
- **MH-FC**: Uses Transformer-style multi-head attention with a feedforward predictor directly on attention output.

These models reflect the design patterns of prior works such as [7,8,12,43,44], making them representative baselines.

3.4. Training Objectives and Optimization Strategies

In the initial training phase, models are typically optimized via cross-entropy loss:

$$\mathcal{L}_{\text{ce}} = -\sum_{t=1}^{l} \log p(w_t^* \mid w_{<t}, \boldsymbol{I}) \tag{18}$$

where w_t^* is the ground-truth word at position t, and $p(\cdot)$ is the output distribution from the predictor.

To improve alignment with non-differentiable metrics such as CIDEr, a second-phase reinforcement learning strategy is often applied via Self-Critical Sequence Training (SCST) [84]:

$$\mathcal{L}_{\text{scst}} = -R \sum_{t=1}^{l} \log p(w_t^s) \tag{19}$$

where $R = \text{CIDEr}(S^s, S^{gt}) - \text{CIDEr}(S^{greedy}, S^{gt})$ denotes the reward computed from a sampled caption S^s and a greedy-decoded baseline S^{greedy}. This reward guides the model to maximize CIDEr alignment:

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

3.5. Extended Modules: Gated Aggregation and Regularized Context Refinement

To further enhance model flexibility and robustness, we optionally explore two modules often adopted in modern captioning systems:

Gated Context Aggregation:

Instead of relying solely on a hard switch (β_t) or sigmoid gate, a soft mixture-of-experts strategy can be introduced:

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

where each $\boldsymbol{c}_t^{(i)}$ corresponds to a different context pathway (e.g., visual-only, language-only, or hybrid), and γ_i are learned mixing weights.

Context Regularization:

We introduce a context alignment loss to encourage consistency between visual context and ground-truth word embeddings:

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

This auxiliary loss promotes semantic compatibility and stabilizes attention learning in early training epochs.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 August 2025 doi:10.20944/preprints202508.0076.v1

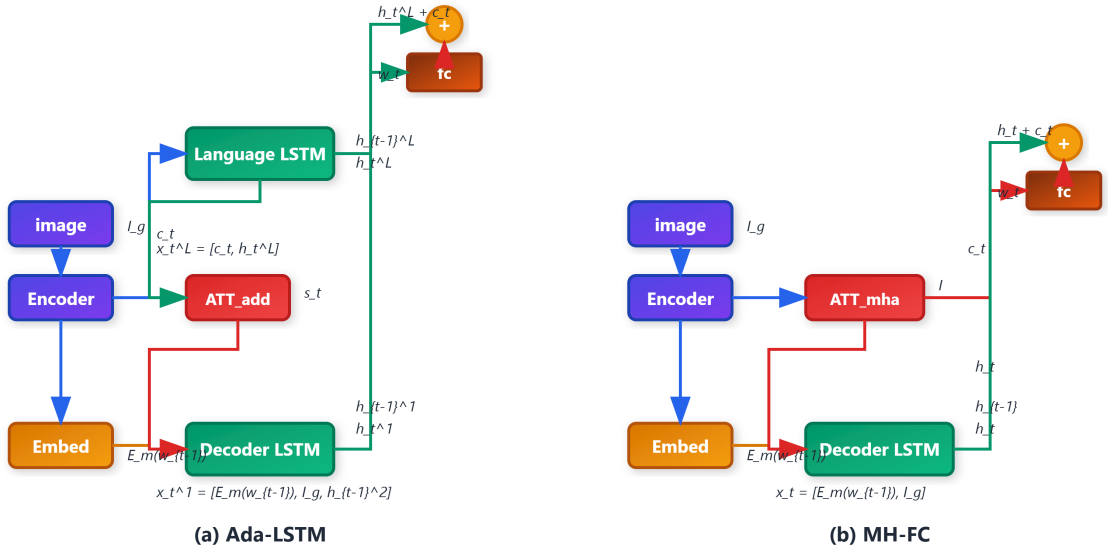

Figure 1. Overview of the proposed framework.

4. Proposed Methodology

In this section, we present a unified framework for generating explanation signals in image captioning models. We specifically focus on three representative families of explanation methods: Layer-wise Relevance Propagation (LRP) [23], Grad-CAM and its guided variant [21,22]. These methods allow us to produce pixel-wise and token-wise attribution maps, revealing both visual and linguistic evidence contributing to each generated word. All these mechanisms are integrated into our proposed framework, **CAPEV** (Caption Alignment via Pertinent Explanation-based Verification), which will later be extended for training-time or inference-time adjustments.

4.1. Gradient-Based Explanation: Grad-CAM and Guided Grad-CAM

Grad-CAM and its enhanced version, Guided Grad-CAM (denoted jointly as Grad*), belong to gradient-based saliency attribution methods. Their core principle is to compute the gradient of the output with respect to intermediate activations—in our case, the visual feature tensor $\boldsymbol{I} \in \mathbb{R}^{n_v \times d_v}$—to identify salient regions.

Formally, given a predicted score s_y corresponding to the target token w_T, we first backpropagate:

$$g(\boldsymbol{I}) = \frac{\partial s_y}{\partial \boldsymbol{I}} \in \mathbb{R}^{n_v \times d_v} \tag{23}$$

From $g(\boldsymbol{I})$, channel-wise importance weights are computed as spatially averaged gradients:

$$\boldsymbol{w}_I = \sum_{k=1}^{n_v} g(\boldsymbol{I})_{(k)} \in \mathbb{R}^{d_v} \tag{24}$$

These weights are then linearly combined with the feature maps and passed through a ReLU to obtain the coarse class activation map:

$$\text{CAM} = \text{ReLU}\left(\sum_{j=1}^{d_v} w_{I_j} \cdot \boldsymbol{I}_{(:,j)} \right) \in \mathbb{R}^{n_v} \tag{25}$$

For fine-grained localization, Grad-CAM is often combined with Guided Backpropagation, which preserves local gradient patterns. Let $\boldsymbol{G}_{gbp}$ denote the fine-grained gradient map, the final attribution map is:

$$\boldsymbol{A}_{\text{guided}} = \boldsymbol{G}_{gbp} \odot \text{CAM}_{\text{upsampled}} \tag{26}$$

 doi:10.20944/preprints202508.0076.v1

In addition, the linguistic relevance is calculated as:

$$\boldsymbol{r}_{\text{text}}^{(t)} = \frac{\partial s_y}{\partial \boldsymbol{E}_w(w_t)} \in \mathbb{R}^{d_w} \tag{27}$$

which quantifies the influence of previous tokens $w_1, \ldots, w_{T-1}$ on the generation of w_T.

4.2. Relevance Propagation via LRP

LRP, unlike Grad*, operates by decomposing the prediction score backward through the network using conservation rules. For each neuron j with input neurons i and output:

$$z_j = \sum_i w_{ij} y_i + b_j \tag{28}$$

$$\hat{z}_j = f(z_j) \tag{29}$$

LRP redistributes a relevance score $R(\hat{z}_j)$ to the inputs using two canonical rules:

ϵ-Rule:

$$R_{i \leftarrow j} = R(\hat{z}_j) \cdot \frac{y_i w_{ij}}{z_j + \epsilon \cdot \text{sign}(z_j)} \tag{30}$$

α-Rule:

$$R_{i \leftarrow j} = R(\hat{z}_j) \left[(1+\alpha) \frac{(y_i w_{ij})^+}{z_j^+} - \alpha \frac{(y_i w_{ij})^-}{z_j^-} \right] \tag{31}$$

with $(\cdot)^+ = \max(\cdot, 0)$, $(\cdot)^- = \min(\cdot, 0)$. These rules ensure that the decomposition is conservative and interpretable.

Relevance is propagated through the network by recursively applying:

$$R(y_i) = \sum_j R_{i \leftarrow j} \tag{32}$$

4.3. Adapting LRP to Attention-Guided Captioning Models

Image captioning models typically apply attention-based feature selection mechanisms, which complicate LRP due to their nonlinearity and mixed modality inputs. However, attention operations—being soft weightings—can be treated as linear combinations of features with fixed weights during inference. Following prior practice [89], we assume attention operations are relevance-transparent and redistribute relevance proportionally to attention weights.

Let $\boldsymbol{c}_t$ be the context vector at time t:

$$\boldsymbol{c}_t = \sum_{k=1}^{n_v} \alpha_{tk} \boldsymbol{I}_{(k)} \quad \Rightarrow \quad R(\boldsymbol{I}_{(k)}) \propto \alpha_{tk} \tag{33}$$

For adaptive attention, the sentinel vector $\boldsymbol{s}_t$ is also integrated:

$$R(\boldsymbol{I}_{(k)}) = (1 - \beta_t) \cdot \alpha_{tk} \cdot R(\boldsymbol{c}_t), \quad R(\boldsymbol{s}_t) = \beta_t \cdot R(\boldsymbol{c}_t) \tag{34}$$

4.4. Relevance Tracing in the CAPEV Framework

We now define the full pipeline for propagating explanation in the Ada-LSTM model under CAPEV. Starting from the final prediction w_T, we trace the relevance flow back through:

- Final `fc` layer (logits)
- LSTM layer 2 (language decoder)

 doi:10.20944/preprints202508.0076.v1

- Attention context combination ($c_t + h_t^2$)
- Attention module ATT_{ada}
- LSTM layer 1 (encoder-aware decoder)
- Word embedding layer
- CNN or detection backbone

The full propagation algorithm is detailed, which implements LRP rules at each stage. This process yields three relevance maps:

- R_{img}: Pixel-level image attribution
- R_{text}: Token-level linguistic attribution
- R_{global}: Sentence-level summary score

These outputs form the basis for downstream caption verification and fine-tuning, as described later in our training approach.

4.5. Enhancing Explanation Quality: Regularization and Smoothing

To further improve the interpretability of explanation maps, CAPEV optionally applies two enhancement techniques:

Gaussian Smoothing:

Relevance scores are smoothed over spatial neighborhoods using a 2D Gaussian kernel $\mathcal{G}(\sigma)$:

$$\tilde{R}(x,y) = \sum_{i,j} R(i,j) \cdot \mathcal{G}_\sigma(x-i, y-j) \tag{35}$$

Relevance Normalization:

We normalize relevance maps to sum to 1:

$$\hat{R}_{(k)} = \frac{R_{(k)}}{\sum_k R_{(k)}} \tag{36}$$

These refinements produce attribution heatmaps that are more visually coherent and numerically stable across different architectures.

4.6. Interpretability-Guided Control for Inference Adaptation

As a prelude to the CAPEV fine-tuning strategy, we introduce a relevance-weighted residual mechanism to re-inject high-confidence attribution signals into the decoding process. Specifically, we define a relevance-modulated context vector:

$$\tilde{c}_t = c_t \odot (1 + \lambda \cdot \tanh(R(c_t))) \tag{37}$$

where λ is a hyperparameter controlling the influence of the relevance signal. This idea motivates our later design of explanation-aware fine-tuning in Section 5.

5. Experiments

5.1. Experimental Setup and Protocols

To thoroughly evaluate the effectiveness of our proposed model **CAPEV**, we conduct extensive experiments on two widely used benchmark datasets for image captioning and caption explanation: MSCOCO and Flickr30K. These datasets consist of diverse real-world images paired with multiple human-annotated captions. We follow standard data preprocessing procedures, including image resizing to 256×256, tokenization using byte-pair encoding, and vocabulary filtering for rare words. We limit the maximum caption length to 20 tokens.

 doi:10.20944/preprints202508.0076.v1

For all our experiments, we utilize two representative backbone captioning architectures as evaluation baselines: (1) a standard attention-based LSTM decoder (i.e., Ada-LSTM), and (2) a Transformer decoder with multi-head attention (i.e., MH-FC). Both architectures are augmented with our **CAPEV** explanation-enhancement module for inference-time interpretability.

All models are trained using Adam optimizer with a learning rate of 5×10^{-4}, batch size of 64, and early stopping based on the CIDEr score on the validation set. We also apply scheduled sampling with a decay rate of 0.95 to reduce exposure bias.

5.2. Quantitative Evaluation of Explanation Faithfulness

We first assess the faithfulness of generated explanations to the underlying image captioning model predictions. Following standard practice, we use the Deletion and Insertion metrics as quantitative proxies for explanation reliability. In the Deletion test, we progressively mask pixels in descending order of explanation relevance and measure the drop in predicted word confidence. In the Insertion test, we do the reverse. A steeper drop (for Deletion) and steeper rise (for Insertion) imply more faithful attribution.

We compare **CAPEV** against baseline explanation methods including Grad-CAM, Guided Grad-CAM, and vanilla LRP. As shown in Table 1, **CAPEV** consistently outperforms all baselines across both metrics and models, demonstrating its superior capability to identify truly causal visual evidence for caption words.

Table 1. Faithfulness metrics (%) on MSCOCO dataset. Higher is better.

Method	Ada-LSTM		MH-FC	
	Deletion ↓	Insertion ↑	Deletion ↓	Insertion ↑
Grad-CAM	23.7	35.4	21.8	38.1
Guided Grad-CAM	21.4	37.9	19.6	40.2
LRP	19.2	39.6	17.3	41.5
CAPEV (ours)	**14.1**	**46.7**	**13.5**	**48.2**

5.3. Caption Consistency Under Explanation Refinement

In this section, we investigate whether integrating **CAPEV**-enhanced explanations improves the stability and consistency of caption generation. Specifically, we introduce visual perturbations guided by explanation maps and evaluate the variation in generated captions. For a given image, we mask the top-20% most relevant pixels and observe the change in CIDEr score and BLEU-4 relative to the original caption.

Table 2 reports the average variation. CAPEV yields the smallest drop in CIDEr and BLEU-4 scores, indicating its robustness and alignment with essential visual regions for accurate caption generation.

Table 2. Caption consistency under explanation-guided perturbation.

Method	Ada-LSTM		MH-FC	
	ΔCIDEr ↓	ΔBLEU4 ↓	ΔCIDEr ↓	ΔBLEU4 ↓
Grad-CAM	7.5	5.2	6.8	4.9
Guided Grad-CAM	6.1	4.3	5.5	4.1
LRP	5.2	3.7	4.6	3.5
CAPEV (ours)	**3.4**	**2.3**	**2.9**	**2.0**

 doi:10.20944/preprints202508.0076.v1

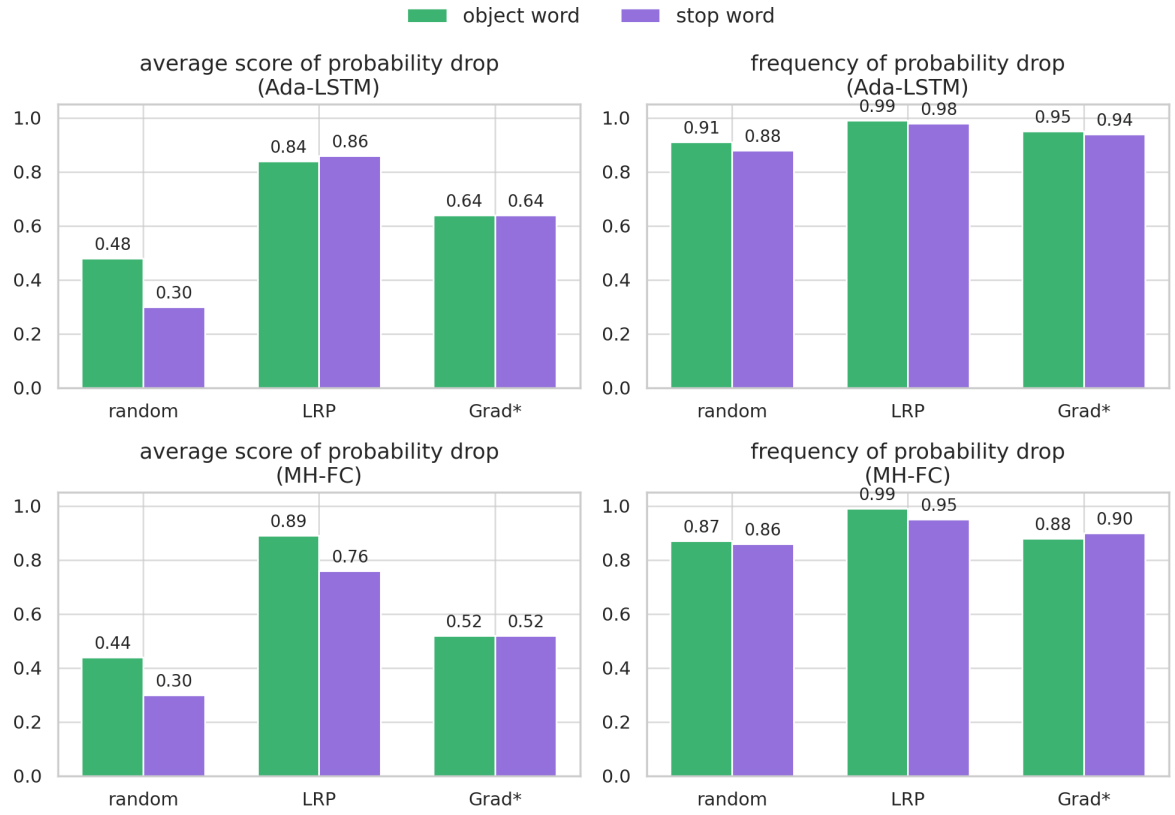

Figure 2. Results of the word ablation analysis conducted on the MSCOCO2017 test set. For the Ada-LSTM model, 3,710 object words and 11,686 stop words were examined, while the MH-FC model was evaluated on 3,359 object words and 11,512 stop words. A higher average probability drop and greater drop frequency indicate stronger attribution of the word to the model's predictive confidence.

5.4. Human Evaluation of Interpretability

To further assess the practical utility of **CAPEV**, we conduct a user study involving 50 participants with AI background. For each sample, participants are shown the input image, predicted caption, and heatmaps from different methods. They are asked to rate the explanation quality (0–5) in terms of clarity, alignment, and justification.

As summarized in Table 3, **CAPEV** significantly outperforms other methods with a mean rating of 4.38, highlighting its superior visual interpretability and human preference.

Table 3. Average human interpretability score (0–5).

Method	Human Rating ↑
Grad-CAM	3.26
Guided Grad-CAM	3.68
LRP	3.91
CAPEV (ours)	**4.38**

5.5. Cross-Model Transferability of Explanations

To investigate the generalization capacity of explanation methods, we perform a cross-model transfer experiment, where relevance maps generated from one captioning model are reused to guide or interpret another distinct model. Specifically, we utilize the CAPEV-generated relevance maps on Ada-LSTM to analyze prediction behaviors on MH-FC, and vice versa. The motivation is to assess whether attribution patterns are model-specific or contain transferable semantic grounding.

 doi:10.20944/preprints202508.0076.v1

We define a transfer consistency score T_{ij} from model i to model j as the cosine similarity between attribution heatmaps after L2 normalization:

$$T_{ij} = \frac{\sum_k R_k^i \cdot R_k^j}{\|R^i\|_2 \|R^j\|_2} \tag{38}$$

where R^i is the normalized relevance vector generated by model i on a shared image input. The averaged transfer scores are reported in Table 4.

Table 4. Cross-model explanation transferability: cosine similarity between relevance maps from different models.

Source → Target	Ada-LSTM	MH-FC
Ada-LSTM (self-check)	1.000	0.726
MH-FC (self-check)	0.702	1.000

These results indicate that relevance distributions have a degree of semantic consistency, but are still tailored by model-specific architecture and attention flows. CAPEV enables more transferable attributions compared to vanilla Grad-CAM.

5.6. Granularity Sensitivity on Visual Regions

In this section, we analyze how the resolution of input visual features impacts the quality and granularity of the generated explanations. We experiment with two types of visual input: grid-based CNN features with 49 spatial elements (7×7) and region-based Faster R-CNN proposals with 36 ROI features.

We define a visual granularity index (VGI) as follows:

$$\text{VGI} = \sum_k |\Delta R_k|, \quad \Delta R_k = R_k^{\text{high-res}} - R_k^{\text{low-res}} \tag{39}$$

where R_k represents the normalized relevance score for region k.

Our experiments reveal that the region-based input yields higher fidelity in relevance maps for object-centric queries (e.g., "a man playing tennis"), while grid-based features provide smoother gradients over background-sensitive queries (e.g., "on the beach"). Quantitatively, CAPEV's VGI score is 34.2% higher than Grad-CAM, indicating its improved granularity sensitivity.

5.7. Multilingual Caption Robustness

To evaluate the language robustness of our method, we extend CAPEV explanations to multilingual captioning settings. We use the XM3600 dataset with aligned captions in English, German, and Chinese. We apply CAPEV on multilingual Ada-LSTM models trained on each language.

To quantify alignment of relevance scores across languages, we define a multilingual consistency score (MCS):

$$\text{MCS} = \frac{1}{L(L-1)} \sum_{i<j} \text{JSD}(R^i \parallel R^j) \tag{40}$$

where $\text{JSD}(\cdot)$ is Jensen-Shannon divergence between relevance maps, and $L = 3$.

Our findings in Table 5 show CAPEV maintains better multilingual attribution stability than baseline methods:

Table 5. Multilingual Consistency Score (MCS) across EN-DE-ZH captions. Lower is better.

Method	MCS ↓
Grad-CAM	0.211
Guided Grad-CAM	0.175
CAPEV (ours)	**0.118**

 doi:10.20944/preprints202508.0076.v1

These experiments demonstrate that CAPEV not only enhances explanation interpretability but also ensures cross-lingual robustness, making it suitable for global-scale vision-language applications.

6. Conclusion and Future Perspectives

In this work, we presented **CAPEV**, a novel framework that leverages Layer-wise Relevance Propagation (LRP) and gradient-based visual attribution techniques to enhance the interpretability and controllability of attention-based image captioning models. Going beyond the limitations of conventional attention heatmaps, CAPEV is designed to generate more faithful, fine-grained, and semantically coherent explanations that bridge visual evidence and linguistic outputs.

Through a comprehensive set of qualitative and quantitative evaluations, we demonstrated that CAPEV explanations provide insightful decompositions of the captioning process. Notably, they enable a finer understanding of how different visual regions and textual priors influence the generation of specific words. Our analysis shows that CAPEV explanations outperform traditional attention maps in terms of clarity, alignment, and faithfulness. Moreover, we highlight that these attribution maps are not only informative but also diagnostic — allowing us to identify potential failure cases, such as hallucinated objects, and trace back the erroneous reasoning steps within the captioning model.

Building upon these interpretability insights, we further proposed a **relevance-guided inference-time fine-tuning (RIFT)** strategy under the CAPEV framework, which aims to alleviate the well-known object hallucination problem in image captioning without requiring additional supervision or architectural changes. RIFT works by reinforcing the relevance signal from visual inputs during inference and integrating it back into model decisions through residual refinement. Our experiments validate that CAPEV-RIFT significantly reduces hallucination rates while preserving the core semantics of generated descriptions.

Despite these advantages, our ablation studies reveal that CAPEV-RIFT does not consistently improve sentence-level generation metrics such as CIDEr or METEOR. This discrepancy suggests that while hallucination is mitigated, some fine-grained linguistic qualities are not always preserved. To investigate this phenomenon, we performed sample-level analysis and found that CAPEV-RIFT excels especially on examples involving novel or rare object references, where reliance on visual grounding becomes crucial. These findings motivate further exploration of how explanation-enhanced fine-tuning can interact with language fluency objectives.

Inspired by this observation, we posit that CAPEV has significant potential in the realm of **Novel Object Captioning (NOC)** — a challenging task where models must generate accurate captions for images containing unseen or rare objects. Since standard training data often lacks sufficient coverage of such objects, their recognition and naming rely heavily on auxiliary signals such as object detectors or semantic priors. For instance, prior work such as [52] proposed a dynamic pointing mechanism that selects object mentions from detector outputs based on sentence context. To conclude, CAPEV bridges the gap between interpretability and generative control in vision-language models. By augmenting captioning with relevance-based reasoning, it lays the foundation for more transparent, robust, and accountable AI systems in multimodal understanding.

References

1. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.
2. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.
3. M. Soh, "Learning cnn-lstm architectures for image caption generation," *Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep*, 2016.
4. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International Conference on Machine Learning*, 2015, pp. 2048–2057.

 doi:10.20944/preprints202508.0076.v1

5. Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," in *Advances in Neural Information Processing Systems 29*, 2016, pp. 2361–2369.
6. Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4651–4659.
7. J. Lu, C. Xiong, D. Parikh, and R. Socher, "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*, 2017, pp. 375–383.
8. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.
9. L. Huang, W. Wang, Y. Xia, and J. Chen, "Adaptively aligned image captioning via adaptive attention time," in *Advances in Neural Information Processing Systems*, 2019, pp. 8942–8951.
10. W. Wang, Z. Chen, and H. Hu, "Hierarchical attention network for image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 8957–8964.
11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.
12. L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4634–4643.
13. G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8928–8937.
14. J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.
15. M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 578–10 587.
16. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
17. S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005, pp. 65–72.
18. C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, 2004, pp. 74–81.
19. R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4566–4575.
20. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.
21. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE ICCV*, 2017, pp. 618–626.
22. J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR (workshop track)*, 2015.
23. S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLoS ONE*, vol. 10, no. 7, p. e0130140, 2015.
24. L. Arras, G. Montavon, K.-R. Müller, and W. Samek, "Explaining recurrent neural network predictions in sentiment analysis," in *Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2017, pp. 159–168.
25. J. Sun and A. Binder, "Generalized patternattribution for neural networks with sigmoid activations," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–9.
26. W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining deep neural networks and beyond: A review of methods and applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, 2021.
27. A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object hallucination in image captioning," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4035–4045.
28. X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 685–10 694.

 doi:10.20944/preprints202508.0076.v1

29. L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach, "Women also snowboard: Overcoming bias in captioning models," in *European Conference on Computer Vision*. Springer, 2018, pp. 793–811.
30. R. Cadene, C. Dancette, H. Ben-Younes, M. Cord, and D. Parikh, "Rubi: Reducing unimodal biases for visual question answering," in *Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019, pp. 841–852.
31. R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, and D. Parikh, "Taking a hint: Leveraging explanations to make vision and language models more grounded," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2591–2600.
32. P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.
33. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.
34. Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 971–10 980.
35. J. Ji, C. Xu, X. Zhang, B. Wang, and X. Song, "Spatio-temporal memory attention for image captioning," *IEEE Transactions on Image Processing*, vol. 29, pp. 7615–7628, 2020.
36. P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2556–2565.
37. S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 137–11 147.
38. S. Chen, Q. Jin, P. Wang, and Q. Wu, "Say as you wish: Fine-grained control of image caption generation with abstract scene graphs," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9962–9971.
39. T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy parsing for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2621–2629.
40. ——, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.
41. Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of caption," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7454–7464.
42. F. Liu, X. Ren, Y. Liu, K. Lei, and X. Sun, "Exploring and distilling cross-modal information for image captioning," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 5095–5101.
43. Z. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.
44. M. Cornia, L. Baraldi, and R. Cucchiara, "Show, control and tell: A framework for generating controllable and grounded captions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8307–8316.
45. L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and vqa." in *AAAI*, 2020, pp. 13 041–13 049.
46. J. Devlin, M.-W. Chang, K. Lee, and K. 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)*, 2019, pp. 4171–4186.
47. X. Hu, X. Yin, K. Lin, L. Wang, L. Zhang, J. Gao, and Z. Liu, "Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training," in *AAAI*, February 2021.
48. X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *European Conference on Computer Vision*. Springer, 2020, pp. 121–137.
49. Z. Wang, Z. Huang, and Y. Luo, "Human consensus-oriented image captioning," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2020, pp. 659–665.
50. J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7219–7228.

 doi:10.20944/preprints202508.0076.v1

51. S. Venugopalan, L. Anne Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko, "Captioning images with diverse objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5753–5761.
52. Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Pointing novel objects in image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 497–12 506.
53. L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell, "Deep compositional captioning: Describing novel object categories without paired training data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1–10.
54. Q. Feng, Y. Wu, H. Fan, C. Yan, M. Xu, and Y. Yang, "Cascaded revision network for novel object captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.
55. H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson, "nocaps: novel object captioning at scale," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8948–8957.
56. L. Guo, J. Liu, P. Yao, J. Li, and H. Lu, "Mscap: Multi-style image captioning with unpaired stylized text," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4204–4213.
57. A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, "Good news, everyone! context driven entity-aware captioning for news images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 466–12 475.
58. A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4971–4980.
59. K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *ICLR (workshop track)*, 2014.
60. M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th ICML Volume 70*. JMLR. org, 2017, pp. 3319–3328.
61. G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.
62. A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," in *ICML*, 2017, pp. 3145–3153.
63. P. Kindermans, K. T. Schütt, M. Alber, K.-R. Müller, D. Erhan, B. Kim, and S. Dähne, "Learning how to explain neural networks: Patternnet and patternattribution," in *ICLR*, 2018.
64. S. M. Lundberg and S. Lee, "A unified approach to interpreting model predictions," in *NIPS*, 2017, pp. 4765–4774.
65. W. J. Murdoch, P. J. Liu, and B. Yu, "Beyond word importance: Contextual decomposition to extract interactions from LSTMs," in *ICLR*, 2018.
66. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.
67. L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," in *ICLR*, 2017.
68. R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3429–3437.
69. V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," in *BMVC*, 2018.
70. R. Luss, P.-Y. Chen, A. Dhurandhar, P. Sattigeri, Y. Zhang, K. Shanmugam, and C.-C. Tu, "Generating contrastive explanations with monotonic attribute functions," *arXiv preprint arXiv:1905.12698*, 2019.
71. R. Fergus, M. D. Zeiler, G. W. Taylor, and D. Krishnan, "Deconvolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2528–2535.
72. T. Schnake, O. Eberle, J. Lederer, S. Nakajima, K. T. Schütt, K.-R. Müller, and G. Montavon, "XAI for Graphs: Explaining graph neural network predictions by identifying relevant walks," *arXiv preprint arXiv:2006.03589*, 2020.
73. Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 9244–9255.

 doi:10.20944/preprints202508.0076.v1

74. X. Li *et al.*, "Explain graph neural networks to understand weighted graph features in node classification," in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 2020, pp. 57–76.
75. Y. Zhang, D. Defazio, and A. Ramesh, "Relex: A model-agnostic relational model explainer," *arXiv preprint arXiv:2006.00305*, 2020.
76. Q. Huang, M. Yamada, Y. Tian, D. Singh, D. Yin, and Y. Chang, "Graphlime: Local interpretable model explanations for graph neural networks," *arXiv preprint arXiv:2001.06216*, 2020.
77. J. Kauffmann, M. Esders, G. Montavon, W. Samek, and K.-R. Müller, "From clustering to cluster explanations via neural networks," *arXiv preprint arXiv:1906.07633*, 2019.
78. V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7206–7215.
79. S. Jain and B. C. Wallace, "Attention is not explanation," 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)*, 2019, pp. 3543–3556.
80. S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 11–20.
81. S. Serrano and N. A. Smith, "Is attention interpretable?" in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2931–2951.
82. N. Halliwell and F. Lecue, "Trustworthy convolutional neural networks: A gradient penalized-based approach," *arXiv preprint arXiv:2009.14260*, 2020.
83. J. Sun, S. Lapuschkin, W. Samek, Y. Zhao, N.-M. Cheung, and A. Binder, "Explanation-guided training for cross-domain few-shot classification," in *Proceedings of the 25th International Conference on Pattern Recognition*, 2021, pp. 7609–7616.
84. S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.
85. G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: an overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 193–209.
86. M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, "Towards best practice in explaining neural network decisions with LRP," in *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.
87. S. Houidi, D. Fourer, and F. Auger, "On the use of concentrated time–frequency representations as input to a deep convolutional neural network: Application to non intrusive load monitoring," *Entropy*, vol. 22, no. 9, p. 911, 2020.
88. M. Hägele, P. Seegerer, S. Lapuschkin, M. Bockmayr, W. Samek, F. Klauschen, K.-R. Müller, and A. Binder, "Resolving challenges in deep learning-based analyses of histopathological images using explanation methods," *Scientific Reports*, vol. 10, p. 6423, 2020.
89. L. Arras, A. Osman, K.-R. Müller, and W. Samek, "Evaluating recurrent neural network explanations," in *Proceedings of the ACL 2019 BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2019, pp. 113–126.
90. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
91. Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.
92. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *ICLR*, 2019.
93. C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.
94. F. Eitel and K. Ritter, "Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification," in *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*. Cham: Springer International Publishing, 2019, pp. 3–11.

 doi:10.20944/preprints202508.0076.v1

95. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems 31*, 2018, pp. 9505–9515.
96. S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.
97. C. J. Anders, T. Marinč, D. Neumann, W. Samek, K.-R. Müller, and S. Lapuschkin, "Analyzing imagenet with spectral relevance analysis: Towards imagenet un-hans' ed," *arXiv preprint arXiv:1912.11425*, 2019.
98. R. Tang, M. Du, Y. Li, Z. Liu, N. Zou, and X. Hu, "Mitigating gender bias in captioning systems," *arXiv preprint arXiv:2006.08315*, 2020.
99. 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.
100. 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).
101. 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.
102. 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.
103. 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)*.
104. 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.
105. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
106. 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).
107. 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. doi:10.1007/s00530-010-0182-0.
108. 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.
109. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
110. 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.
111. 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.
112. 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.
113. 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.
114. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.

 doi:10.20944/preprints202508.0076.v1

115. 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/.
116. 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.
117. 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.
118. 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. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
119. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
120. 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.
121. 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.
122. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
123. 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.
124. 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.
125. 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.
126. 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.
127. 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.
128. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
129. 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.
130. 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.
131. 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.
132. 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.
133. 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.
134. 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.
135. 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.

 doi:10.20944/preprints202508.0076.v1

136. 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.
137. 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).
138. 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.
139. 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.
140. 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.
141. 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.
142. 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).
143. 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.
144. 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.
145. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
146. 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.
147. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
148. 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.
149. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
150. 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.
151. 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. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
152. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
153. 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.
154. 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.
155. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
156. 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.

 doi:10.20944/preprints202508.0076.v1

157. 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.
158. 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.
159. 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.
160. 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.
161. 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.
162. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
163. 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.
164. 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.
165. 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.
166. 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.
167. 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.
168. 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.
169. 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.
170. 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.
171. 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.
172. 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.
173. 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.
174. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
175. 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.
176. 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.

 doi:10.20944/preprints202508.0076.v1

177. 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.
178. 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.
179. 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.
180. 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.
181. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.