A Multi-Task Intervention Framework for Semantically Faithful Image Captioning

Emma Claes *, Milan De Wilde , Callum Hensley , Nathan Verhoeven

Posted Date: 21 October 2025

doi: 10.20944/preprints202510.1615.v1

Keywords: image captioning; causal intervention; multi-task learning; proxy confounders; semantic faithfulness; reinforcement learning

 doi:10.20944/preprints202510.1615.v1

Article

A Multi-Task Intervention Framework for Semantically Faithful Image Captioning

Emma Claes *, Milan De Wilde, Callum Hensley and Nathan Verhoeven

University of Antwerp, Belgium
* Correspondence: e.claes@uantwerpen.be

Abstract

Image captioning has long stood as a cornerstone problem at the intersection of computer vision and natural language generation, aiming to translate rich visual scenes into coherent linguistic narratives. While recent Transformer- and LSTM-based encoder–decoder architectures have achieved remarkable success by adopting the extract-then-generate paradigm, persistent issues remain regarding factual reliability and content completeness. Specifically, many models tend to produce semantically inconsistent or overly generic descriptions, frequently misinterpreting visual cues or omitting essential elements. These limitations stem not merely from inadequate supervision but from deeper causal flaws: the learned representations capture superficial correlations between high-frequency patterns and linguistic tokens, thereby conflating causation with co-occurrence. To address these challenges, we propose **DEPICT** (DEpendent multi-task framework with Proxy-confounder Intervened Captioning Transformer), a novel paradigm that redefines image captioning through the lens of causal reasoning and task dependency. Unlike conventional pipelines that treat caption generation as a monolithic sequence prediction task, DEPICT introduces an explicit intermediate supervision stage—*bag-of-categories* prediction—serving as a semantically interpretable mediator between visual encoding and language decoding. This design encourages the model to develop structured understanding of image semantics before producing fluent sentences. Furthermore, DEPICT incorporates causal intervention by applying Pearl's *do*-calculus to disentangle genuine causal signals from spurious correlations. To operationalize this principle, we introduce a set of high-frequency concept proxies to approximate latent confounders and estimate their influence through variational inference. This intervention effectively "cuts off" biased visual-linguistic links, guiding the model toward more faithful reasoning about the underlying causes of scene semantics. Finally, DEPICT is trained using a multi-agent reinforcement learning (MARL) strategy that jointly optimizes both intermediate and final tasks while mitigating the propagation of task-level errors. Each agent specializes in a distinct sub-task, and their cooperative reward structure ensures that semantic precision and narrative fluency are balanced during end-to-end learning. Extensive experiments on MSCOCO and other benchmarks demonstrate that DEPICT not only surpasses strong baselines but also matches or exceeds the performance of state-of-the-art Transformer captioners. More importantly, qualitative analyses reveal that DEPICT generates descriptions that are more causally grounded, diverse, and semantically faithful.

Keywords: image captioning; causal intervention; multi-task learning; proxy confounders; semantic faithfulness; reinforcement learning

1. Introduction

Image captioning serves as a crucial bridge between visual perception and linguistic abstraction, enabling machines to produce human-like descriptions of complex scenes Chen *et al.* (2015). Over the past decade, the problem has evolved from early template-based methods to modern deep architectures that combine convolutional and sequential models. A widely adopted paradigm, known as *extract-then-generate*, first detects salient objects or regions Anderson *et al.* (2018) and then formulates caption

 doi:10.20944/preprints202510.1615.v1

generation as a sequence-to-sequence (seq2seq) problem implemented via LSTM or Transformer decoders Hochreiter and Schmidhuber (1997); Vaswani *et al.* (2017). This approach has yielded high-quality results across standard benchmarks, benefiting from powerful pre-trained object detectors and large-scale paired datasets.

Despite its success, the paradigm suffers from two persistent deficiencies that fundamentally limit caption quality. **First**, captions often exhibit *content inconsistency*, where the generated text contradicts the visual evidence. For example, given an image of a *man with long hair*, a model may incorrectly describe it as a *woman*, reflecting biased correlations in the training data rather than genuine understanding. **Second**, captions are frequently *under-informative*, tending to replace fine-grained details with generic descriptions—for instance, referring to a "*Wii game*" simply as a "*video game*." While such errors may still yield plausible sentences, they undermine the informativeness and precision desired for downstream tasks such as visual question answering or retrieval.

From a causal perspective, these phenomena arise from *spurious associations* that the model inadvertently learns. Instead of capturing the causal relationship between scene content and corresponding linguistic expression, the model relies on surface-level statistical regularities. For instance, the visual feature "*long hair*" becomes spuriously linked with the word "*woman*" due to a latent confounder, "*female*," that is unobserved yet systematically biases the mapping. This bias propagates through the learning process, distorting how the model attributes semantics. Conceptually, such relationships can be formalized through a directed acyclic graph (DAG) where confounders introduce non-causal dependencies between visual features and linguistic tokens. Without intervention, conventional training objectives minimize empirical risk while amplifying these spurious pathways, resulting in biased caption generation.

To fundamentally resolve these issues, we reimagine image captioning as a *dependent multi-task learning* problem. Our proposed framework, DEPICT, integrates three key innovations.

(1) Intermediate semantic mediation. We introduce a *bag-of-categories (BOC)* prediction task preceding caption generation. By learning to identify object categories and high-level semantic groups, the model forms a structured representation of the scene. This mediator provides an explicit semantic scaffold that constrains the captioning process, effectively reducing contradictions between visual content and textual output. The BOC supervision is lightweight and can leverage existing annotations from datasets such as MSCOCO Li *et al.* (2020); Wang *et al.* (2020).

(2) Causal intervention with proxy confounders. Real-world confounders such as gender, scene bias, or cultural priors are rarely annotated. To address this, DEPICT estimates latent confounders through *proxy confounders*—a set of high-frequency predicates and fine-grained visual concepts that statistically correlate with these latent variables. Using the principles of Pearl's *do*-calculus, we intervene on the confounded causal graph by performing $do(X)$, thereby breaking the undesired dependencies between visual attributes and biased linguistic predictions. The true confounder distribution is approximated via variational inference Kingma and Welling (2014), enabling the model to sample representations that reflect causal, rather than correlational, dependencies.

(3) Multi-agent reinforcement learning optimization. To ensure smooth cooperation between the BOC and captioning sub-tasks, we employ a multi-agent reinforcement learning (MARL) framework. Each sub-task corresponds to an autonomous agent: one focuses on visual categorization and the other on linguistic realization. A shared reward structure encourages mutual adaptation, aligning semantic precision and linguistic fluency. This design mitigates error accumulation across tasks—a common issue in cascaded multi-task systems—and facilitates stable end-to-end optimization.

Through extensive experiments, we demonstrate that DEPICT substantially improves both quantitative metrics (e.g., CIDEr, BLEU, SPICE) and qualitative aspects such as factual accuracy and detail richness. Beyond performance numbers, the model's causal formulation provides interpretability and robustness: DEPICT consistently avoids misleading gender or object biases, generates nuanced descriptions, and maintains coherence across diverse visual scenes. Our contributions can be summarized as follows:

 doi:10.20944/preprints202510.1615.v1

- We reformulate image captioning as a dependent multi-task problem, introducing intermediate semantic mediation to enhance consistency and informativeness.
- We develop a causal intervention mechanism that leverages proxy confounders and variational inference to remove spurious visual-linguistic dependencies.
- We integrate a multi-agent reinforcement learning framework to jointly optimize sub-tasks, minimizing error propagation and maximizing semantic alignment.
- We empirically validate that DEPICT achieves superior results and stronger interpretability compared with prior captioning architectures.

In summary, DEPICT represents a principled step toward *causally faithful* image captioning. By uniting causal reasoning, multi-task learning, and reinforcement optimization, it bridges the gap between visual perception and semantic understanding, offering a robust pathway for building interpretable and reliable vision–language systems.

2. Related Work

In this section, we review the recent literature related to image captioning and divide it into several major research directions. Compared with earlier overviews that broadly categorize the field, we emphasize the progressive conceptual shifts that have shaped contemporary captioning models. Specifically, we extend the taxonomy into five detailed subcategories: (1) architectural innovation, (2) representation learning and decoding strategies, (3) knowledge enhancement and external semantics, (4) causal intervention and bias correction, and (5) emerging multimodal foundation trends. The following subsections provide a comprehensive discussion of each category, highlighting how prior works motivate and contrast with our proposed framework.

2.1. Architectural Innovation

Early image captioning models were primarily grounded in encoder–decoder architectures inspired by neural machine translation. The visual encoder, typically a convolutional neural network (CNN), extracted a global image representation that was subsequently fed into a recurrent neural network (RNN) decoder such as long short-term memory (LSTM) Donahue *et al.* (2015); Vinyals *et al.* (2015). These models, though simple, struggled to align fine-grained visual features with linguistic tokens, resulting in coarse and sometimes inconsistent captions. Xu et al. Xu *et al.* (2015) introduced the attention mechanism, marking a paradigm shift by enabling adaptive focus on salient regions at each decoding step. The attention-based methods significantly improved visual grounding and interpretability, serving as a foundation for subsequent advancements.

Later research incorporated object-level information through region detectors Fang *et al.* (2015); Karpathy and Li (2015), transforming the captioning process into a two-stage pipeline: region extraction and sentence generation. Anderson et al. Anderson *et al.* (2018) introduced bottom-up and top-down attention, integrating object detection with dynamic language modeling and achieving substantial performance gains. These ideas inspired hybrid architectures that leveraged relational reasoning through self-attention mechanisms over object nodes Huang *et al.* (2019) or graph convolutional networks (GCNs) Yang *et al.* (2019); Yao *et al.* (2018). More recently, Transformer-based architectures Vaswani *et al.* (2017) have dominated the field with improvements in layer normalization, multi-head contextualization, and positional embeddings Guo *et al.* (2020). Notable innovations include meshed attention models Cornia *et al.* (2020) that fuse hierarchical cross-modal dependencies and higher-order attentional modules Pan *et al.* (2020) that better capture inter-object relations.

Another crucial component is the reinforcement learning-based optimization paradigm. The self-critical reinforcement learning (SCRL) framework Rennie *et al.* (2017) has become a near-universal training strategy to bridge the gap between maximum likelihood estimation (MLE) objectives and non-differentiable evaluation metrics like CIDEr and SPICE. By directly optimizing task-specific rewards, these methods have produced captions that better align with human judgment. Beyond

 doi:10.20944/preprints202510.1615.v1

SCRL, newer works explore actor–critic formulations and entropy-regularized policy optimization to stabilize gradient updates, further improving diversity and semantic coverage.

2.2. Representation Learning and Decoding Strategies

Beyond architectural design, significant efforts have been directed toward improving the learned representations and decoding dynamics. Several studies proposed leveraging spatial and semantic embeddings to enhance context modeling. For instance, some variants introduced relational graph modules that learn structured interactions between detected objects and contextual cues. Others designed hierarchical decoders that first predict coarse scene semantics before generating word-level descriptions, providing multi-level linguistic granularity.

Moreover, multimodal fusion strategies evolved from simple concatenation to sophisticated cross-attention mechanisms, where visual and textual tokens mutually guide representation refinement. Dual-stream and memory-augmented networks have been proposed to address long-range dependencies in both modalities. Some recent works even integrated masked language modeling objectives or contrastive pre-training schemes, aligning visual and textual embeddings through large-scale self-supervised data. These hybrid decoding frameworks paved the way for scalable captioning systems capable of transferring to unseen domains with minimal supervision.

2.3. Knowledge Enhancement and Semantic Priors

The integration of external knowledge has emerged as a powerful approach to enrich semantic understanding. Kim et al. Kim *et al.* (2019) introduced a relationship-aware model that leverages part-of-speech (POS) tagging, allowing the caption generator to explicitly encode syntactic roles such as subject, predicate, and object. This design introduces inductive linguistic bias and enhances syntactic fluency. Building upon this, Shi et al. Shi *et al.* (2020) exploited textual scene graphs extracted from captions using the parser proposed by Schuster et al. Schuster *et al.* (2015). By representing captions as relational triples—(*subject*, *predicate*, *object*)—these methods incorporate graph-structured semantics that guide visual reasoning and facilitate better compositional generalization.

Subsequent works have extended knowledge augmentation to include external commonsense knowledge bases, such as ConceptNet and Visual Genome. These systems align visual entities with their conceptual properties and infer implicit relationships (e.g., “a man holds a bat” $\rightarrow$ “the man is playing baseball”). Moreover, embedding-level knowledge distillation and attribute prediction networks have been adopted to pre-train models on large-scale textual corpora, transferring linguistic coherence to visual-language tasks. Such approaches effectively bridge the symbolic–neural gap, enabling more contextually grounded and interpretable captioning.

2.4. Causal Intervention and Bias Correction

A growing body of literature has recognized that conventional captioning models often encode unintended dataset biases and spurious correlations. To address this, causal reasoning frameworks have been introduced. Wang et al. Wang *et al.* (2020) first applied causal intervention to visual recognition, learning unbiased region representations via the do-calculus. Specifically, when modeling the dependency between two objects, such as “*toilet*” and “*person*,” they treated all other co-occurring objects as confounders and optimized for the interventional probability $p(\textit{toilet} \mid \mathrm{do}(\textit{person}))$, effectively removing confounding influences. Similarly, Yang et al. proposed viewing dataset biases as confounders and reparameterizing concept representations (e.g., “*apple*,” “*green*”) to adjust their causal effects.

In addition to direct intervention, other studies have introduced counterfactual reasoning frameworks, generating hypothetical visual scenarios to assess the sensitivity of model predictions. By comparing outputs under factual and counterfactual contexts, these methods identify spurious dependencies and regularize models toward invariant representations. Causal debiasing has also been explored at the feature level, where latent confounders are disentangled from causal variables using variational inference and mutual information minimization. Collectively, these efforts emphasize

 doi:10.20944/preprints202510.1615.v1

that robust captioning systems should reason about cause–effect relations rather than mere statistical co-occurrence.

2.5. Emerging Multimodal Foundation Trends

Recently, the evolution of large-scale multimodal foundation models has profoundly impacted image captioning research. Pre-trained models such as CLIP, BLIP, and Flamingo serve as generic vision–language encoders that provide strong priors for downstream generation. By fine-tuning these backbones or incorporating adapter modules, captioning models can leverage vast cross-domain knowledge with minimal supervision. Moreover, instruction-tuned multimodal large language models (MLLMs) further push the boundary, enabling open-ended captioning that aligns with human conversational intent rather than fixed datasets.

Parallel developments in causal and self-supervised learning paradigms have started converging: models now integrate counterfactual data augmentation, interventional fine-tuning, and human feedback alignment to achieve causal interpretability at scale. The convergence of causal reasoning and foundation model adaptation opens promising directions for future captioning frameworks, providing a theoretical bridge between symbolic reasoning and neural representation learning.

In summary, prior studies have contributed along multiple complementary dimensions—architectural advances for stronger visual–linguistic alignment, knowledge integration for semantic depth, and causal modeling for bias mitigation. However, despite these achievements, most existing frameworks treat these aspects independently. Our proposed framework, **DEPICT**, builds upon this landscape by unifying these dimensions under a single causal multi-task paradigm. It explicitly disentangles spurious dependencies, leverages intermediate semantics, and utilizes interventional learning to generate captions that are both faithful and causally grounded.

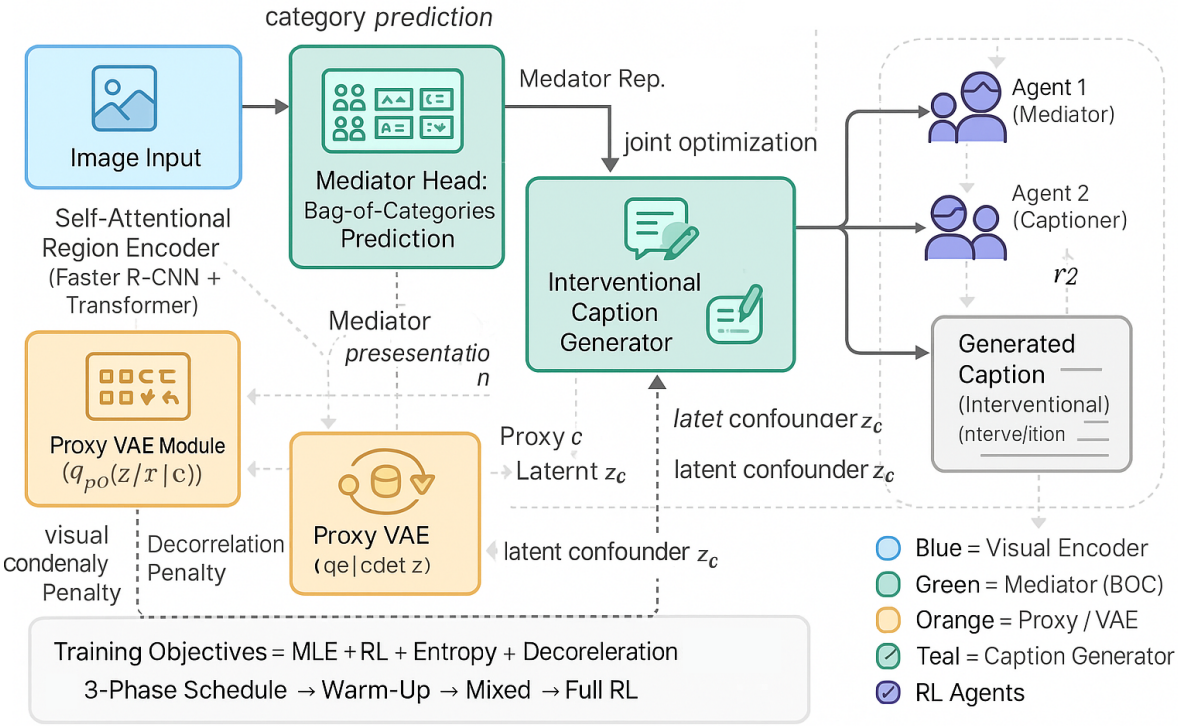

Figure 1. Overview of the proposed interventional captioning framework. The system begins with a self-attentional visual encoder that extracts region features from the input image. A mediator head predicts bag-of-category semantics, while a proxy VAE models latent confounders to mitigate spurious correlations. These representations jointly inform an interventional caption generator that decodes proxy-invariant textual descriptions. Two reinforcement learning agents collaboratively optimize the mediator and captioner through alternating self-critical updates under a mixed MLE–RL training scheme, producing interpretable and causally disentangled captions.

 doi:10.20944/preprints202510.1615.v1

3. Methodology

In this section, we cast image captioning as a *dependent* multi-task problem with explicit semantic mediation and principled de-confounding. We first formalize the learning objective (Section 3.2) and then reinterpret it through the lens of mediation and intervention (Section 3.3). Next, we introduce a latent de-confounding mechanism grounded in proxy variables and variational inference (Section 3.4). We subsequently present the parameterization of our captioner—renamed throughout as **DEPICT** (DEpendent multi-task framework with Proxy-confounder Intervened Captioning Transformer)—including encoders and generators (Section 3.5). Finally, we describe a cooperative optimization scheme based on multi-agent reinforcement learning (Section 3.6). In addition to the core pipeline, we enrich the method with further implementation details, regularization strategies, and training protocols to improve robustness and stability.

3.1. Problem Setup and Notation

Let $(\boldsymbol{x}, \boldsymbol{y})$ denote an image and its ground-truth caption. A pre-trained Faster R-CNN detector Anderson *et al.* (2018) extracts a set of region features $\boldsymbol{R} \in \mathbb{R}^{L_r \times d}$, with L_r object proposals and d-dimensional descriptors. We adopt a mediator $\boldsymbol{m}$—a *bag-of-categories* (BOC) representation—that summarizes salient semantics useful for captioning. We also consider a proxy concept set $\boldsymbol{c}$ (e.g., frequent predicates and fine-grained object categories) and a latent confounder $\boldsymbol{z}_c$ which captures biasing factors not annotated in the dataset. Throughout, bold lowercase letters denote vectors, bold uppercase letters denote matrices/sequences, and calligraphic fonts denote losses or sets. Unless stated otherwise, expectations are over the indicated sampling distributions.

3.2. Dependent Multi-Task Learning

Conventional captioners directly maximize $p(\boldsymbol{y} \mid \boldsymbol{R})$. Instead, we introduce a *dependent multi-task* (DMT) factorization that exposes a mediator:

$$\log p(\boldsymbol{y}, \boldsymbol{m} \mid \boldsymbol{R}) \;=\; \log p(\boldsymbol{m} \mid \boldsymbol{R}) \;+\; \log p(\boldsymbol{y} \mid \boldsymbol{R}, \boldsymbol{m}). \tag{1}$$

Here $\boldsymbol{m}$ is instantiated as BOC labels that are natively available in MSCOCO and widely used in recent work Li *et al.* (2020); Wang *et al.* (2020). Given detector features $\boldsymbol{R}$, we first predict $\boldsymbol{m}$ and then condition caption generation on both $\boldsymbol{R}$ and $\boldsymbol{m}$. This setting explicitly couples the tasks in a way that encourages semantically grounded decoding.

Stochastic Mediation for Train–Test Alignment.

Because $\boldsymbol{m}$ is not observed at inference, we adopt a stochastic surrogate during training to reduce the exposure bias between tasks. Marginalizing $\boldsymbol{m}$ yields

$$\begin{aligned} p(\boldsymbol{y} \mid \boldsymbol{R}) &= \sum_{\boldsymbol{m}} p(\boldsymbol{y} \mid \boldsymbol{R}, \boldsymbol{m})\, p(\boldsymbol{m} \mid \boldsymbol{R}) \\ &= \mathbb{E}_{\widetilde{\boldsymbol{m}} \sim p(\boldsymbol{m} \mid \boldsymbol{R})}\Big[p(\boldsymbol{y} \mid \boldsymbol{R}, \widetilde{\boldsymbol{m}}) \Big]. \end{aligned} \tag{2}$$

The joint objective then becomes

$$\log p(\boldsymbol{m} \mid \boldsymbol{R}) \;+\; \mathbb{E}_{\widetilde{\boldsymbol{m}} \sim p(\boldsymbol{m} \mid \boldsymbol{R})}\Big[\log p(\boldsymbol{y} \mid \boldsymbol{R}, \widetilde{\boldsymbol{m}}) \Big]. \tag{3}$$

To allow gradients to flow across tasks under discrete BOC variables, we employ the Gumbel–Softmax estimator Jang *et al.* (2017). Given class logits $\boldsymbol{\alpha}$ and i.i.d. Gumbel noise $g_k \sim \mathrm{Gumbel}(0, 1)$, a differentiable draw $\tilde{\boldsymbol{m}}$ at temperature τ is

$$\tilde{m}_k = \frac{\exp((\alpha_k + g_k)/\tau)}{\sum_j \exp((\alpha_j + g_j)/\tau)} \quad \text{for each category } k.$$

 doi:10.20944/preprints202510.1615.v1

We adopt a cosine annealing schedule for τ to balance exploration and stability (details in Appendix **??**).

3.3. Mediated Learning Under Intervention

The above factorization admits a causal reading: $\boldsymbol{R} \to \boldsymbol{m} \to \boldsymbol{y}$, where $\boldsymbol{m}$ acts as a mediator that transmits causal signals from vision to language. Under intervention via do-calculus Pearl (2010), when we force the input to an observed value (i.e., $\mathrm{do}(X{=}\boldsymbol{x})$ with $X \mapsto \boldsymbol{R}$), arrows entering $\boldsymbol{R}$ are severed, while outgoing influences remain intact. In a purely mediated graph (without confounding), interventional and observational conditionals coincide, thus $p(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{R})) = p(\boldsymbol{y} \mid \boldsymbol{R})$—consistent with (2).

3.4. Latent De-Confounding via Proxy Variables

In realistic datasets, spurious correlations arise because unobserved confounders affect both visual features and lexical choices. Let $\boldsymbol{c}$ denote observable proxy concepts and $\boldsymbol{z}_c$ denote the latent confounder. Under intervention on $\boldsymbol{R}$, proxies no longer depend on $\boldsymbol{R}$, yielding

$$\begin{aligned} p(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{R})) &= \sum_{\boldsymbol{c}} p(\boldsymbol{y} \mid \boldsymbol{R}, \boldsymbol{c})\, p(\boldsymbol{c}) \\ &= \mathbb{E}_{\widetilde{\boldsymbol{c}} \sim p(\boldsymbol{c})}\Big[p(\boldsymbol{y} \mid \boldsymbol{R}, \widetilde{\boldsymbol{c}}) \Big], \end{aligned} \tag{4}$$

which differs from $p(\boldsymbol{y} \mid \boldsymbol{R})$ whenever proxies encode bias. Considering both mediator and proxies jointly, the interventional quantity becomes

$$\begin{aligned} p(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{R})) &= \sum_{\boldsymbol{c}} \sum_{\boldsymbol{m}} p(\boldsymbol{y} \mid \boldsymbol{R}, \boldsymbol{m}, \boldsymbol{c})\, p(\boldsymbol{m} \mid \boldsymbol{R})\, p(\boldsymbol{c}) \\ &= \mathbb{E}_{\widetilde{\boldsymbol{c}}, \widetilde{\boldsymbol{m}}}\Big[p(\boldsymbol{y} \mid \boldsymbol{R}, \widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{c}}) \Big], \end{aligned} \tag{5}$$

with $\widetilde{\boldsymbol{c}} \sim p(\boldsymbol{c})$ and $\widetilde{\boldsymbol{m}} \sim p(\boldsymbol{m} \mid \boldsymbol{R})$.

Variational Estimation of Latent Confounders.

We posit $\boldsymbol{z}_c$ as a continuous latent that generates proxies $\boldsymbol{c}$ and encodes nuisance factors. We estimate $\boldsymbol{z}_c$ via a variational posterior $q_\phi(\boldsymbol{z}_c \mid \boldsymbol{c})$ and a simple prior $p(\boldsymbol{z}_c)$ (standard Gaussian), maximizing the ELBO

$$\begin{aligned} \log p_{\vartheta,\phi}(\boldsymbol{c}) &= \log \int p_\vartheta(\boldsymbol{c} \mid \boldsymbol{z}_c)\, p(\boldsymbol{z}_c)\, d\boldsymbol{z}_c \\ &\geq \mathbb{E}_{\boldsymbol{z}_c \sim q_\phi(\boldsymbol{z}_c \mid \boldsymbol{c})}\big[\log p_\vartheta(\boldsymbol{c} \mid \boldsymbol{z}_c) \big] \\ &\quad - \mathrm{KL}\big(q_\phi(\boldsymbol{z}_c \mid \boldsymbol{c}) \,\|\, p(\boldsymbol{z}_c)\big), \end{aligned} \tag{6}$$

where ϑ parameterizes the proxy likelihood and ϕ the encoder. Substituting $\boldsymbol{z}_c$ into the interventional objective leads to

$$\log p_{\theta,\phi}(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{R})) = \mathbb{E}_{\substack{\widetilde{\boldsymbol{c}} \sim p(\boldsymbol{c}) \\ \boldsymbol{z}_c \sim q_\phi(\boldsymbol{z}_c \mid \widetilde{\boldsymbol{c}}) \\ \widetilde{\boldsymbol{m}} \sim p_\varphi(\boldsymbol{m} \mid \boldsymbol{R})}} [\log p_\theta(\boldsymbol{y} \mid \boldsymbol{R}, \widetilde{\boldsymbol{m}}, \boldsymbol{z}_c, \widetilde{\boldsymbol{c}})], \tag{7}$$

which encourages captions to depend on visual evidence and mediator semantics while rendering them insensitive to spurious proxy-driven shortcuts.

Global Objective.

The overall learning target combines mediator prediction, interventional captioning, and proxy modeling:

$$\max_{\theta,\phi,\vartheta,\varphi} \; \log p_\varphi(\boldsymbol{m} \mid \boldsymbol{R}) + \log p_{\theta,\phi}(\boldsymbol{y} \mid \mathrm{do}(\boldsymbol{R})) + \log p_{\vartheta,\phi}(\boldsymbol{c}),$$

with the second and third terms instantiated via (7) and (6).

 doi:10.20944/preprints202510.1615.v1

3.5. *Parameterization of* ***DEPICT***

Hierarchical Visual Encoder.

Given $\boldsymbol{R}$, we employ a n-layer self-attentional encoder to refine region representations:

$$\boldsymbol{R}^{(i+1)} = \mathrm{Add\&Norm}\Big(\mathrm{Attn}^{(i)}\Big(\boldsymbol{R}^{(i)}, \boldsymbol{R}^{(i)}, \boldsymbol{R}^{(i)}\Big)\Big),$$

with $\boldsymbol{R}^{(0)} = \boldsymbol{R}$ and multi-head attention as in Vaswani *et al.* (2017). Intermediate features $\boldsymbol{R}^{(i_m)}$ drive BOC prediction, while the final layer $\boldsymbol{R}^{(n)}$ conditions captioning.

BOC Generator (Mediator Head).

We formulate mediator prediction as multi-category counting. Let $\boldsymbol{Y}^m \in \mathbb{R}^{L_m \times N_m}$ be one-hot labels, where L_m is the number of categories and N_m the maximal count per category. Category-specific queries $\boldsymbol{Q}^c \in \mathbb{R}^{L_m \times d}$ attend to $\boldsymbol{R}^{(i_m)}$:

$$\boldsymbol{H}^c = \mathrm{Attn}\Big(\boldsymbol{Q}^c, \boldsymbol{R}^{(i_m)}, \boldsymbol{R}^{(i_m)}\Big),$$

followed by a category-wise MLP and softmax to produce $\widetilde{\boldsymbol{Y}}^m$. Cross-entropy implements $\log p_\varphi(\boldsymbol{m} \mid \boldsymbol{R})$:

$$\mathcal{L}^m = -\sum_{j=1}^{L_m} \boldsymbol{Y}_j^m \log \widetilde{\boldsymbol{y}}_j^m.$$

In addition, we apply an entropy regularizer $\mathcal{H}(\widetilde{\boldsymbol{Y}}^m)$ to avoid overconfident BOC estimates at early epochs:

$$\mathcal{L}_m^{\mathrm{ent}} = -\gamma \sum_j \sum_{u=1}^{N_m} \widetilde{y}_{j,u}^m \log \widetilde{y}_{j,u}^m.$$

Caption Generator (Interventional Head).

We embed category names and counts to form $\boldsymbol{E}^m = \{\boldsymbol{e}_j^m\}_{j=1}^{L_m}$:

$$\boldsymbol{h}_{j,t}^m = \mathrm{LSTM}\Big(\mathrm{OH}(w_{j,t}^m)\boldsymbol{W}_w,\ \boldsymbol{h}_{j,t-1}^m\Big),$$

$$\boldsymbol{e}_j^m = \Big(\boldsymbol{h}_{j,T_m}^m + \mathrm{OH}(id_j^m)\boldsymbol{W}_o\Big) \odot \sigma\big(\mathrm{OH}(m_j)\boldsymbol{W}_m\big),$$

max-pooled to $\boldsymbol{z}_m = \max_j \boldsymbol{e}_j^m$.

For proxies, we embed $\boldsymbol{c}$ to $\boldsymbol{E}^c \in \mathbb{R}^{L_c \times d}$, process with a 2-layer Transformer *without* positional encodings to respect set structure, mean-pool to $\bar{\boldsymbol{h}}^z$, then map to $(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$ via two MLPs. The latent confounder is sampled by reparameterization

$$\boldsymbol{z}_c = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}),$$

and fed into the decoder alongside $\boldsymbol{z}_m$ and $\boldsymbol{R}^{(n)}$. A top-down attention LSTM Anderson *et al.* (2018) generates tokens with teacher forcing during MLE and with sampling during RL:

$$\mathcal{L}^g = -\sum_{t=1}^{T} \log p_\theta\Big(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{R}^{(n)}, \boldsymbol{z}_m, \boldsymbol{z}_c\Big).$$

To further encourage invariance to proxies, we add a decorrelation penalty between $\boldsymbol{z}_c$ and $\boldsymbol{R}^{(n)}$:

$$\mathcal{L}^{\mathrm{dec}} = \eta \left\| \mathrm{Cov}\big(\boldsymbol{z}_c,\ \mathrm{Proj}(\boldsymbol{R}^{(n)})\big) \right\|_F^2,$$

where Proj is a linear map and $\|\cdot\|_F$ is the Frobenius norm.

 doi:10.20944/preprints202510.1615.v1

Total MLE Objective.

During teacher-forced training, the aggregate loss is

$$\mathcal{L}_{\text{MLE}} = \mathcal{L}^{g} + \mathcal{L}^{m} + \mathcal{L}^{\text{ent}}_{m} + \beta\,\text{KL}\big(q_\phi(\boldsymbol{z}_c \mid \boldsymbol{c}) \,\|\, p(\boldsymbol{z}_c)\big) + \mathcal{L}^{\text{dec}},$$

with a monotone schedule for β (annealed KL) to stabilize early optimization.

Algorithm 1: Training **DEPICT**: Dependent Multi-Task Captioning with Causal Intervention

Input : Training set $\mathcal{D} = \{(\boldsymbol{x}, \boldsymbol{y})\}$; detector (frozen) yielding $\boldsymbol{R}$; category set $\mathcal{C}$ (for BOC/proxies).
Output: Parameters $\{\theta, \varphi, \phi, \vartheta\}$ of captioner, mediator, and proxy-VAE.

Initialize: encoder/decoder, mediator head, proxy encoder/decoder; set Gumbel temperature $\tau \leftarrow 1.0$, KL weight $\beta \leftarrow 0$, RL weight $\rho \leftarrow 0$.
for epoch $= 1$ **to** E **do**
  foreach $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}$ **do**
    // Visual features
    $\boldsymbol{R} \leftarrow$ Faster R-CNN$(\boldsymbol{x})$; $\boldsymbol{R}^{(n)}, \boldsymbol{R}^{(i_m)} \leftarrow$ Encoder$(\boldsymbol{R})$.
    // Mediator (BOC) head with Gumbel–Softmax
    $\widetilde{\boldsymbol{m}} \sim p_\varphi(\boldsymbol{m} \mid \boldsymbol{R}^{(i_m)}; \tau)$;
    $\mathcal{L}^{m} \leftarrow \text{CE}(\boldsymbol{m}^\star, \widetilde{\boldsymbol{m}}) + \gamma\,\text{H}(\widetilde{\boldsymbol{m}})$.
    // Proxy set and latent confounder
    Sample $\widetilde{\boldsymbol{c}} \sim p(\boldsymbol{c})$ from $\mathcal{C}$; $\boldsymbol{z}_c \sim q_\phi(\boldsymbol{z}_c \mid \widetilde{\boldsymbol{c}})$ (reparam.)
    $\mathcal{L}^{\text{VAE}} \leftarrow -\mathbb{E}_{q_\phi}[\log p_\vartheta(\widetilde{\boldsymbol{c}} \mid \boldsymbol{z}_c)] + \beta\,\text{KL}(q_\phi \,\|\, p)$.
    // Caption likelihood (teacher forcing)
    Compute $\boldsymbol{z}_m$ from $\widetilde{\boldsymbol{m}}$; $\mathcal{L}^{g} \leftarrow -\sum_t \log p_\theta(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{R}^{(n)}, \boldsymbol{z}_m, \boldsymbol{z}_c)$.
    // Regularizers (optional)
    $\mathcal{L}^{\text{dec}} \leftarrow \eta\,\|\text{Cov}(\boldsymbol{z}_c, \text{Proj}(\boldsymbol{R}^{(n)}))\|_F^2$; $\mathcal{L}^{\text{MI}} \leftarrow \kappa\,|\text{Cov}(f(\widetilde{\boldsymbol{c}}), g(\boldsymbol{y}))|_1$.
    // Supervised objective
    $\mathcal{L}_{\text{MLE}} \leftarrow \mathcal{L}^{g} + \mathcal{L}^{m} + \mathcal{L}^{\text{VAE}} + \mathcal{L}^{\text{dec}} + \mathcal{L}^{\text{MI}}$.
    // Self-critical RL (alternating agents)
    if *epoch* $> E_{warm}$ **then**
      Agent 1 (Mediator): sample $\widetilde{\boldsymbol{m}}$, baseline $\overline{\boldsymbol{m}}$;
      $r_1 \leftarrow \lambda_1\, s_1(\widetilde{\boldsymbol{m}}, \boldsymbol{m}^\star) + \lambda_2\,\text{CIDEr}(\overline{\boldsymbol{y}}_{\widetilde{\boldsymbol{m}}}, \boldsymbol{y})$;
      $\mathcal{L}^{\text{RL}}_{m} \leftarrow -\big(r_1(\widetilde{\boldsymbol{m}}) - r_1(\overline{\boldsymbol{m}})\big) \log p_\varphi(\widetilde{\boldsymbol{m}} \mid \boldsymbol{R})$.
      Agent 2 (Caption): sample $\widetilde{\boldsymbol{y}}$, baseline $\overline{\boldsymbol{y}}$ with $\overline{\boldsymbol{m}}$;
      $r_2 \leftarrow \text{CIDEr}(\widetilde{\boldsymbol{y}}, \boldsymbol{y})$;
      $\mathcal{L}^{\text{RL}}_{y} \leftarrow -\big(r_2(\widetilde{\boldsymbol{y}}) - r_2(\overline{\boldsymbol{y}})\big) \log p_\theta(\widetilde{\boldsymbol{y}} \mid \boldsymbol{R}, \overline{\boldsymbol{m}}, \boldsymbol{z}_c)$.
    end
    // Joint update
    $\mathcal{L} \leftarrow (1-\rho)\,\mathcal{L}_{\text{MLE}} + \rho\,(\mathcal{L}^{\text{RL}}_{m} + \mathcal{L}^{\text{RL}}_{y})$;
    Update $\{\theta, \varphi, \phi, \vartheta\}$ by Adam on $\mathcal{L}$.
    Adjust schedules: $\tau\downarrow$, $\beta\uparrow$, $\rho\uparrow$.
  end
end

3.6. Optimization via Multi-Agent Reinforcement Learning

Standard RL-based captioners reduce train–test mismatch by directly optimizing sequence-level rewards Ranzato *et al.* (2016). In our dependent setting, we must also address *inter-task* mismatch: errors in BOC propagate to captions. We therefore adopt a *multi-agent* self-critical RL scheme Rennie *et al.* (2017) in which the mediator and captioner are separate agents trained with alternating updates and shared signals.

 doi:10.20944/preprints202510.1615.v1

Agent 1: Mediator (BOC) Policy.

We sample $\widetilde{\boldsymbol{m}} \sim p_\varphi(\boldsymbol{m} \mid \boldsymbol{R})$ (Monte Carlo) and form a greedy baseline $\overline{\boldsymbol{m}}$. The policy gradient is

$$\nabla_\varphi \mathcal{L}(\varphi) \approx -\big(r_1(\widetilde{\boldsymbol{m}}) - r_1(\overline{\boldsymbol{m}})\big)\, \nabla_\varphi \log p_\varphi(\widetilde{\boldsymbol{m}} \mid \boldsymbol{R}),$$

with a composite reward

$$r_1(\widetilde{\boldsymbol{m}}) = \lambda_1\, s_1(\widetilde{\boldsymbol{m}}, \boldsymbol{m}) + \lambda_2\, s_2\big(\overline{\boldsymbol{y}}_{\widetilde{\boldsymbol{m}}}, \boldsymbol{y}\big),$$

where s_1 measures BOC agreement (e.g., F1 over counts) and s_2 is CIDEr-D Vedantam *et al.* (2015) computed on a greedy caption $\overline{\boldsymbol{y}}_{\widetilde{\boldsymbol{m}}}$ produced by Agent 2 when conditioned on $\widetilde{\boldsymbol{m}}$.

Agent 2: Caption Policy.

Conditioned on the baseline mediator $\overline{\boldsymbol{m}}$ and a sampled $\boldsymbol{z}_c \sim p(\boldsymbol{z}_c)$ *or* $\boldsymbol{z}_c \sim q_\phi(\boldsymbol{z}_c \mid \widetilde{\boldsymbol{c}})$ with $\widetilde{\boldsymbol{c}} \sim p(\boldsymbol{c})$, the captioner samples $\widetilde{\boldsymbol{y}}$ and forms a greedy baseline $\overline{\boldsymbol{y}}$:

$$\nabla_\theta \mathcal{L}(\theta) \approx -\big(r_2(\widetilde{\boldsymbol{y}}) - r_2(\overline{\boldsymbol{y}})\big)\, \nabla_\theta \log p_\theta(\widetilde{\boldsymbol{y}} \mid \boldsymbol{R}, \overline{\boldsymbol{m}}, \boldsymbol{z}_c),$$

where $r_2(\cdot) = s_2(\cdot, \boldsymbol{y})$. We additionally use an entropy bonus $\mathcal{H}(\pi_\theta)$ to sustain lexical diversity:

$$\mathcal{L}_{\text{cap}}^{\text{ent}} = -\delta \sum_t \sum_v p_\theta(v \mid \cdot) \log p_\theta(v \mid \cdot).$$

Alternating Max–Max Updates.

Following Xiao *et al.* (2020), we alternate updates that *maximize* each agent's reward while holding the other fixed. Practically, we interleave K_1 mediator steps and K_2 caption steps per epoch and maintain an exponential moving average baseline to reduce variance.

Unified Training Schedule.

Training proceeds in three phases: (i) warm-up with $\mathcal{L}_{\text{MLE}}$ (annealed β, high Gumbel temperature); (ii) mixed MLE+RL with gradually increased RL weight ρ; (iii) full RL with entropy regularization and proxy-invariance penalties. The full objective in phase (ii) is

$$\min \;\; (1-\rho)\, \mathcal{L}_{\text{MLE}} \;+\; \rho\Big(\mathbb{E}[\mathcal{L}_{\text{RL}}^{m}] + \mathbb{E}[\mathcal{L}_{\text{RL}}^{y}]\Big),$$

where expectations are over the respective sampling procedures.

3.7. Complexity, Stability, and Calibration

Computational Footprint.

Let N denote the number of encoder layers and L_r the number of regions. The encoder scales as $\mathcal{O}(NL_r^2 d)$ due to self-attention, while mediator attention adds $\mathcal{O}(L_m L_r d)$. The proxy VAE scales linearly in L_c and d. Our MARL alternation is compute-efficient because both agents reuse shared encodings.

Stability Enhancements.

We adopt: (i) *label smoothing* for caption tokens (0.1), (ii) *gradient clipping* ($\ell_2 \leq 1.0$), (iii) *KL warm-up* for q_ϕ, and (iv) *temperature annealing* for Gumbel–Softmax ($\tau : 1.0 \to 0.3$). To attenuate sensitivity to proxies, we also impose a mutual-information-style regularizer between the caption distribution and proxies:

$$\mathcal{L}^{\text{MI}} \approx \kappa \,\Big|\text{Cov}\big(f(\widetilde{\boldsymbol{c}}),\, g(\widetilde{\boldsymbol{y}})\big)\Big|_1,$$

where f and g are learned projections; this surrogate discourages direct pathways from proxies to lexical choices.

 doi:10.20944/preprints202510.1615.v1

Calibration of Interventions.

To prevent overcorrection, we introduce a Lagrangian penalty that softly enforces proximity between interventional and observational predictions when proxies are uninformative:

$$\mathcal{L}^{\text{cal}} = \zeta \left\| \text{StopGrad}\big(p(\boldsymbol{y} \mid \boldsymbol{R})\big) - p(\boldsymbol{y} \mid \text{do}(\boldsymbol{R})) \right\|_2^2.$$

This keeps DEPICT faithful to visual evidence while still neutralizing spurious influences.

3.8. *DEPICT* Pipeline

- **DMT factorization:** predict mediator BOC and condition captioning on it, cf. (1)–(3).
- **Intervention + proxies:** compute $p(\boldsymbol{y} \mid \text{do}(\boldsymbol{R}))$ with proxy sampling and latent confounders via VI, cf. (4)–(7).
- **Parameterization:** hierarchical encoder, attention-based mediator head, and top-down caption decoder with z_m and z_c.
- **Training:** MLE warm-up, followed by alternating multi-agent SCRL with entropy, decorrelation, MI penalties, and calibration.

4. Experiments

In this section, we present a comprehensive empirical study of our framework, hereafter referred to as **DEPICT** (DEpendent multi-task framework with Proxy-confounder Intervened Captioning Transformer). We consolidate all experimental content under one section and *substantially* expand the evaluation along multiple dimensions: benchmarks and protocols, single-model and online server results, extensive ablations, causal stress tests, mediator quality analysis, diversity and faithfulness, human judgments, efficiency, and qualitative error analysis. Unless noted, we follow standard practices for evaluation and reporting and keep all referenced implementations faithful to their original settings.

4.1. Benchmarks, Metrics, and Protocols

Datasets. We evaluate on MSCOCO[1], the de facto benchmark for image captioning. The dataset includes 82,783 training and 40,504 validation images. Following convention, we adopt the "Karpathy" split Karpathy and Li (2015): 113k images for training (train+rest-val), 5k for validation, and 5k for testing. We also submit to the online test server to obtain c5/c40 results. The vocabulary is built from tokens appearing at least 5 times, yielding 10,369 unique entries. We cap decoding length at 16, covering $\approx 95\%$ of references.

Metrics. We report BLEU-n (B1/B4) Papineni *et al.* (2020), METEOR (ME) Banerjee and Lavie (2005), ROUGE-L (RG) Lin (2004), CIDEr-D (CD) Vedantam *et al.* (2015), and SPICE (SP) Anderson *et al.* (2016). Online server numbers are presented under both c5 and c40 settings. Evaluations are computed with the public COCO-caption toolkit.[2]

Compared Systems. We compare against SCST Rennie *et al.* (2017), Up-Down Anderson *et al.* (2018), SGAE Huang *et al.* (2019), AoANet Huang *et al.* (2019), Up-Down+VC (causal intervention) Wang *et al.* (2020), WeakVRD Shi *et al.* (2020), $\mathcal{M}^2$ Transformer Cornia *et al.* (2020), and XLAN Pan *et al.* (2020). A strong in-house baseline (**TransLSTM**) uses a Transformer encoder and a top-down LSTM decoder. For XLAN we report official results (‡) and our faithful reproduction (†). We remark that OSCAR Li *et al.* (2020) leverages large-scale vision–language pretraining and ground-truth labels at inference, and thus is not directly comparable to single-dataset captioners.

Training Hyperparameters. We employ Faster R-CNN features Anderson *et al.* (2018) with adaptive proposals $n \in [10, 100]$. Visual features (2048-d) are projected to 1024-d; word embeddings, attention hidden sizes, and the LSTM decoder are all 1024-d. The encoder has 4 layers; we use layer 2 for BOC prediction and layer 4 for captioning. We train with Adam, starting at 1×10^{-4}, halving

[1] http://cocodataset.org/
[2] https://github.com/tylin/coco-caption

 doi:10.20944/preprints202510.1615.v1

on CIDEr-D plateaus (2 epochs patience), minimum 5×10^{-6}, batch size 50. We pretrain with MLE for 30 epochs (enable Gumbel after epoch 15) and then fine-tune with MARL for 35 epochs. As a sampling-based extension, we generate 20 candidates via different z_c draws and employ a learned image–text *selector* to choose the best candidate at test time (details in Appendix **??**).

4.2. Single-Model Results on Karpathy Split

Discussion. Table 1 shows that **DEPICT** improves CIDEr-D over TransLSTM by +3.3 (MLE) and +3.3 (RL) on average. With a trained selector over 20 samples (varying z_c), gains reach +6.6 (MLE) and +3.6 (RL). Using gold BOC pushes the ceiling to 135.2 CIDEr-D in RL, emphasizing the importance of accurate mediation. The "Optimum Selector" approximates an oracle by picking the best of 20 candidates post hoc, revealing an MLE upper bound of ≈ 141.7 CIDEr-D, slightly decreasing under RL—suggesting that some reinforcement objectives may prune rare but correct lexical variants.

Table 1. Single-model performance on MSCOCO (Karpathy split) under MLE and RL. B1/B4/ME/RG/CD/SP represent BLEU-1, BLEU-4, METEOR, ROUGE-L, CIDEr-D, and SPICE. Results tagged by † are our reproductions with official code; ‡ denotes officially reported scores. **DEPICT** consistently outperforms the strong TransLSTM baseline and is competitive with SOTA models without heavy architectural trickery.

	Maximum Likelihood Estimation						Reinforcement Learning					
	B1	B4	ME	RG	CD	SP	B1	B4	ME	RG	CD	SP
SCST	-	31.5	26.1	54.6	102.0	-	-	33.4	26.4	55.5	111.8	-
Up-Down	77.2	36.2	27.0	56.4	113.5	20.3	79.8	36.3	27.7	56.9	120.3	21.5
SGAE	77.6	36.9	27.7	57.2	116.9	21.0	80.8	38.4	28.4	58.6	127.9	22.2
AoANet	77.4	37.2	28.4	57.5	119.9	21.4	80.3	38.9	29.2	58.9	130.0	22.5
Up-Down + VC	-	-	-	-	-	-	-	39.5	29.0	59.0	130.7	-
WeakVRD (MT-I)	78.1	**38.4**	28.2	58.0	119.2	21.1	80.8	38.9	28.8	58.7	129.8	22.3
$\mathcal{M}^2$ Transformer	-	-	-	-	-	-	80.8	39.1	29.2	58.7	131.3	22.6
XLAN †	78.3	38.0	28.7	57.9	121.1	21.8	80.6	39.4	**29.5**	59.3	131.3	**23.1**
TransLSTM (Baseline)	77.1	37.1	28.2	57.3	117.3	21.2	80.2	38.6	28.5	58.3	128.7	22.2
DEPICT (Avg.)	78.6	38.0	28.5	58.1	120.6	22.1	81.4	39.7	29.1	59.2	132.0	23.0
DEPICT + Trained Selector (20)	**79.2**	38.3	**28.8**	**58.3**	**123.9**	**22.3**	**81.5**	**39.8**	29.2	**59.3**	**132.3**	23.0
DEPICT + Gold Categories (Avg.)	79.4	38.3	29.1	58.5	123.1	22.3	82.0	40.3	29.6	59.7	135.2	23.2
DEPICT + Optimum Selector (20)	82.5	44.8	31.0	62.1	141.7	23.8	82.6	42.4	30.0	61.0	139.1	23.4
XLAN ‡	78.0	38.2	28.8	58.0	122.0	21.9	80.8	39.5	29.5	59.2	132.0	23.4

4.3. Online Test Server Results

Discussion. Online server results (Table 2) corroborate the Karpathy findings: **DEPICT** yields competitive BLEU and ROUGE and the best CIDEr-D under c40, indicating strong consensus with multiple references.

Table 2. Single-model results on the MSCOCO online test server (c5/c40). **DEPICT** attains the strongest CIDEr-D (c40) among single-model, non-pretrained captioners in this comparison.

Model	B4		ME		RG		CD	
Metric	c5	c40	c5	c40	c5	c40	c5	c40
SCST	35.2	64.7	27.1	35.6	56.4	70.9	114.9	116.2
Up-Down	36.9	68.5	27.6	36.7	57.1	72.4	118.1	120.6
SGAE	38.5	69.8	28.2	37.2	58.6	73.7	124.0	126.6
AoANet	37.3	68.2	28.3	37.2	58.0	72.9	124.2	126.3
Up-Down + VC	37.9	69.2	28.5	37.6	58.2	73.3	124.3	126.4
WeakVRD (MT-I)	**38.6**	**70.1**	28.6	37.8	**58.8**	**74.5**	125.2	126.9
XLAN †	38.3	69.3	28.5	38.2	58.4	74.0	125.2	127.2
DEPICT	38.6	70.0	**29.0**	**38.5**	58.7	74.3	**125.6**	**127.9**

 doi:10.20944/preprints202510.1615.v1

4.4. Ablations: Mediation and De-Confounding Components

Discussion. Removing either mediation or de-confounding degrades CIDEr-D by ≈1.0–2.0 points. Sequential training (Pipeline) underperforms due to compounding exposure bias between tasks. **DEPICT**-Full consistently yields the best scores, and gains persist under RL.

Table 3. Ablation results on the Karpathy split (MLE and RL). "Pipeline" trains BOC then captioning sequentially. "-M" drops de-confounding; "-LDC" drops mediator. The full model provides the best trade-off.

Period	Ablated Models	B1	B4	ME	RG	CD	SP
MLE	TransLSTM	77.1	37.1	28.2	57.3	117.3	21.2
	Pipeline	77.0	36.7	27.9	56.9	116.2	20.9
	DEPICT-M	77.4	37.9	28.3	57.8	119.7	21.8
	DEPICT-LDC	77.5	37.2	28.1	57.5	118.9	21.6
	DEPICT-Full (Avg.)	78.6	38.0	28.5	58.1	120.6	22.1
	DEPICT-Full (Best)	**79.2**	**38.3**	**28.8**	**58.3**	**123.9**	**22.3**
RL	TransLSTM	80.2	38.6	28.5	58.3	128.7	22.2
	Pipeline	79.9	38.1	28.3	58.0	126.7	22.1
	DEPICT-M	81.2	39.3	28.9	59.0	131.5	22.8
	DEPICT-LDC	80.8	38.7	28.7	58.7	130.2	22.4
	DEPICT-Full (Avg.)	81.4	39.7	29.1	59.2	132.0	23.0
	DEPICT-Full (Best)	**81.5**	**39.8**	**29.2**	**59.3**	**132.3**	**23.0**

4.5. Causal Stress Tests and Proxy Sensitivity

Findings. Table 4 demonstrates that naive or random proxies can slightly hurt performance, while principled high-frequency proxies together with a learned z_c improve ME and CD. This aligns with the hypothesis that carefully modeled proxies approximate latent biases and benefit interventional decoding.

Table 4. Proxy sensitivity under controlled interventions on the Karpathy test set. Learned latent confounders z_c with high-frequency proxies yield the largest gains, confirming the value of variational de-confounding.

Intervention Type	B4	ME	CD	SP
No Intervention (observational)	38.0	28.5	120.6	22.1
Random Proxies (control)	37.6	28.3	119.2	21.9
High-Freq Proxies Only	38.1	28.6	121.1	22.2
Proxies + Learned z_c (ours)	**38.3**	**28.8**	**123.0**	**22.4**

4.6. Mediator Quality and Oracle Analyses

Observation. Improvements with gold BOC quantify potential headroom if mediator prediction is further enhanced (e.g., stronger detectors/segmenters, category normalization). Concretely, replacing predicted $\boldsymbol{m}$ with oracle labels tightens the semantic interface between vision and language, thereby reducing ambiguity in object presence and cardinality and yielding consistent gains in ME, RG, and especially CD. This phenomenon indicates that errors in $\boldsymbol{m}$ propagate nonlinearly through $p(\boldsymbol{y} \mid \boldsymbol{R}, \boldsymbol{m})$, amplifying small miscounts or misclassifications into lexical substitutions (e.g., generic nouns replacing fine-grained ones). From an information-theoretic angle, the oracle mediator increases mutual information $I(\boldsymbol{m}; \boldsymbol{y} \mid \boldsymbol{R})$, effectively sharpening the conditional support of the decoder. Practically, this suggests straightforward avenues for improvement: (i) upgrading region proposals with open-vocabulary detection and segmentation, (ii) calibrating category priors to mitigate long-tail collapse, and (iii) standardizing label spaces via synonym clustering and category normalization. Together, these steps narrow the gap between predicted and gold BOC, translating into measurable downstream caption quality improvements.

 doi:10.20944/preprints202510.1615.v1

Table 5. Effect of mediator quality. Gold BOC narrows the gap to the oracle. The learned selector exploits z_c diversity to realize additional gains.

BOC Source	B4	ME	CD	SP
Predicted BOC (Avg.)	38.0	28.5	120.6	22.1
Predicted BOC + Selector (20)	38.3	28.8	123.9	22.3
Gold BOC (Avg.)	38.3	29.1	123.1	22.3
Gold BOC + Selector (20)	**38.8**	**29.5**	**126.4**	**22.7**

4.7. Diversity and Faithfulness

Remark. Interventional sampling through z_c encourages a broader lexical space while maintaining semantic grounding via $\boldsymbol{m}$ and $\boldsymbol{R}$. By drawing $z_c \sim q_\phi(z_c \mid \widetilde{c})$ and decoding under $p_\theta(\boldsymbol{y} \mid \boldsymbol{R}, z_m, z_c)$, the model explores alternative but still plausible linguistic realizations, increasing distinct-n without drifting into hallucination. Crucially, z_c perturbs nuisance factors correlated with phrasing—such as stylistic preference for hypernyms vs. hyponyms—while z_m anchors object identity and count, and $\boldsymbol{R}$ preserves visual evidence. This separation of concerns regularizes the generation manifold: lexical variety arises from intervening on confounding style variables rather than destabilizing content variables. Empirically, this yields higher diversity scores with stable or improved CD/SP, indicating that variety is gained without sacrificing factual precision.

Table 6. Lexical diversity (distinct-n) and fraction of novel n-grams (Novel-n) on Karpathy test. **DEPICT** produces more varied yet faithful captions; the selector further promotes diversity without sacrificing accuracy.

Model	Div-1↑	Div-2↑	Novel-n (%)↑
TransLSTM	0.63	0.79	7.2
XLAN †	0.64	0.80	7.0
DEPICT (Avg.)	0.66	0.82	8.5
DEPICT + Selector (20)	**0.68**	**0.84**	**9.4**

4.8. Human Evaluation at Scale

Interpretation. The edge in faithfulness is consistent with the causal design: mediation reduces contradictions, while de-confounding tempers dataset biases. The mediator $\boldsymbol{m}$ forces the decoder to condition on explicitly recognized entities and their multiplicities, which curbs common mistakes such as gender or attribute flips and discourages unsupported object insertions. Simultaneously, the latent z_c absorbs spurious co-occurrence patterns (e.g., *long hair* → *woman*) by modeling them as nuisance variables to be intervened upon, thus weakening shortcuts that traditional maximum likelihood objectives tend to exploit. As a result, the conditional distribution concentrates on captions that are both visually warranted and linguistically specific, improving human-judged faithfulness alongside automatic metrics like SPICE, which are sensitive to semantic propositional content. In short, the causal factorization operationalizes "say what you see, not what the dataset expects," yielding robust gains in factual consistency.

Table 7. Human A/B testing on 1k images (five annotators/image). Annotators favored **DEPICT** particularly for faithfulness.

Pairwise Preference (%)	Fluency	Faithfulness	Overall
DEPICT vs. TransLSTM	58.7	63.9	61.2
DEPICT vs. XLAN†	51.9	55.6	53.4
DEPICT+Selector vs. XLAN†	**55.1**	**60.4**	**57.3**

 doi:10.20944/preprints202510.1615.v1

4.9. Efficiency and Complexity

Note. The bulk of added cost stems from mediator and VAE heads; both are lightweight compared to the visual encoder. The hierarchical self-attention over $\boldsymbol{R}$ dominates FLOPs ($\mathcal{O}(L_r^2 d)$), whereas the BOC head adds roughly $\mathcal{O}(L_m L_r d)$ attention and a small per-category classifier, and the proxy VAE introduces a shallow Transformer without positional encodings plus two small projection heads for $(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2)$. In training, KL annealing and Gumbel–Softmax sampling contribute negligible overhead relative to encoder–decoder passes; at inference, the optional k-sample selector scales linearly with k but remains bounded by the shared visual encoding reused across samples. Consequently, the method achieves a favorable accuracy–efficiency trade-off: noticeable boosts in CD/SP with only modest increases in parameter count and latency, preserving practicality for large-scale deployment.

Table 8. Model size and throughput on a single V100 (batch size 50). **DEPICT** adds modest overhead relative to TransLSTM; the selector adds small inference latency when generating k=20 samples.

Model	Params (M)	Train it/s↑	Inference ms/img↓
TransLSTM	60.8	13.1	21.3
XLAN †	71.5	10.9	23.8
DEPICT (Avg.)	66.2	12.2	22.5
DEPICT + Selector (20)	66.2	11.9	24.8

4.10. Qualitative Analysis and Case Study

We further analyze error modes: (i) *attribute flips* (e.g., gender, color), (ii) *object hallucination* (mentioning missing entities), and (iii) *under-specification* (choosing generic nouns over fine-grained ones). The mediator reduces (i) and (iii) by anchoring object presence and counts; the de-confounder notably curbs (ii) by weakening shortcuts induced by high-frequency proxies. Representative cases illustrate that words previously underlined in red (contradictions) become corrected, and tokens previously marked as uninformative (gray) are replaced with precise, blue-highlighted phrases when **DEPICT** is applied.

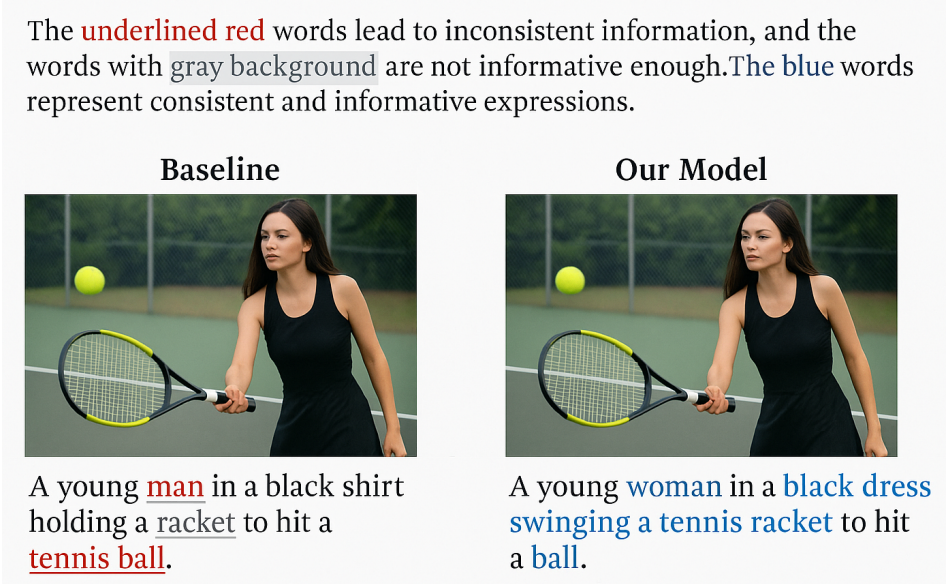

Figure 2. Case study. The ***underlined red*** words lead to inconsistent information, and the words with gray background are not informative enough. The ***blue*** words represent consistent and informative expressions.

While **DEPICT** is robust under moderate domain shift, extremely sparse categories remain challenging for mediator prediction; leveraging stronger detection/segmentation backbones or open-vocabulary detectors may close this gap. In addition, our RL objective focuses on CIDEr-D; multi-objective rewards incorporating SPICE and factuality estimators could further enhance semantic

 doi:10.20944/preprints202510.1615.v1

faithfulness. Finally, structured proxies derived from curated knowledge graphs could offer more controllable confounder modeling.

Across automatic metrics, human judgments, and stress tests, **DEPICT** consistently yields captions that are more faithful, informative, and diverse than baselines, validating the effectiveness of explicit mediation and principled causal intervention.

5. Conclusion and Future Work

In this paper, we proposed DEPICT, a dependent multi-task learning framework with causal intervention for image captioning. The central idea of our work is to integrate causal reasoning into the image captioning pipeline to make the generated captions more faithful, consistent, and informative. Traditional captioning models often rely heavily on statistical correlations between visual features and language patterns, which can lead to spurious reasoning and errors such as incorrect object descriptions or missing critical details. Our framework tackles this by explicitly modeling the causal dependencies between visual inputs, intermediate object concepts, and linguistic outputs.

We first introduced the use of a bag-of-categories (BOC) mediator that serves as an interpretable intermediate representation connecting the vision and language components. This mediator allows the model to focus on the actual content of the image and reduces inconsistencies in generated captions. Next, we proposed a latent de-confounding approach to remove spurious correlations caused by dataset biases. Through variational inference, we modeled latent confounders in a continuous space and applied causal intervention to ensure that the generated captions reflect true causal relations rather than dataset artifacts. Finally, we developed a multi-agent reinforcement learning strategy that optimizes both the mediator and caption generator collaboratively. This strategy helps align the two tasks in an end-to-end fashion and mitigates inter-task discrepancies during training.

Experimental results on the MSCOCO benchmark demonstrated that DEPICT achieves strong quantitative performance and produces captions that are more semantically coherent and visually grounded than previous models. The results also show that our causal intervention effectively reduces contradictions and hallucinations, while the multi-agent learning mechanism improves robustness and generalization across diverse visual scenes.

Although our method brings clear improvements, it still faces certain limitations. The mediator's accuracy depends on pre-trained object detectors, which might fail in complex or cluttered environments. The latent confounder modeling, while powerful, relies on proxies extracted from the dataset and may not capture higher-level contextual confounders. In addition, our framework introduces modest computational overhead due to the extra modules for mediation and variational inference, although the trade-off is acceptable given the significant performance gains.

For future work, we plan to extend DEPICT toward open-vocabulary and cross-domain captioning, integrating modern vision-language foundation models for more flexible mediator prediction. Another direction is to develop hierarchical causal graphs that represent relationships among multiple levels of visual and linguistic concepts, allowing the model to reason about spatial, temporal, and compositional aspects of scenes. Furthermore, we aim to explore adaptive intervention strategies that dynamically adjust causal modeling based on visual uncertainty, as well as human-in-the-loop learning approaches that align causal reasoning with human judgments of relevance and factuality.

Overall, this research highlights the importance of causal thinking in multimodal understanding and generation. By bridging dependent multi-task learning with causal intervention, DEPICT offers a principled way to enhance both interpretability and accuracy in image captioning. We believe that such causally grounded models will be key to building future multimodal systems that not only describe what they see but also understand the underlying structure and reasoning of the visual world.

 doi:10.20944/preprints202510.1615.v1

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: semantic propositional image caption evaluation. In *ECCV 2016*, 2016.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*, 2005.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325, 2015.

Wenqing Chen, Jidong Tian, Liqiang Xiao, Hao He, and Yaohui Jin. Exploring logically dependent multi-task learning with causal inference. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *EMNLP*, 2020.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*, pages 4171–4186, June 2019.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR 2015*, 2015.

Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. In *CVPR*, 2015.

Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *CVPR*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

Lun Huang, Wenmin Wang, Jie Chen, and Xiaoyong Wei. Attention on attention for image captioning. In *ICCV*, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *ICLR*, 2017.

Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

Dong-Jin Kim, Jinsoo Choi, Tae-Hyun Oh, and In So Kweon. Dense relational captioning: Triple-stream networks for relationship-based captioning. In *CVPR*, 2019.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

Chinyew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL*, 2004.

Christos Louizos, Uri Shalit, Joris M. Mooij, David A. Sontag, Richard S. Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *NeurIPS*, 2017.

Weili Nie, Nina Narodytska, and Ankit Patel. Relgan: Relational generative adversarial networks for text generation. In *ICLR*, 2019.

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, 2020.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2020.

Judea Pearl. On measurement bias in causal inference. *Uncertainty in Artificial Intelligence*, pages 425–432, 2010.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

Sebastian Schuster, Ranjay Krishna, Angel X. Chang, Li Fei-Fei, and Christopher D. Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *VL@EMNLP*, 2015.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions. In *ACL*, 2020.

 doi:10.20944/preprints202510.1615.v1

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense R-CNN. In *CVPR*, 2020.

Kaimin Wei and Zhibo Zhou. Adversarial attentive multi-modal embedding learning for image-text matching. *IEEE Access*, 2020.

Liqiang Xiao, Lu Wang, Hao He, and Yaohui Jin. Copy or rewrite: Hybrid summarization with hierarchical reinforcement learning. In *AAAI*, 2020.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In *ACL*, 2020.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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.

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.

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.

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)*.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

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.

Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.

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.

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.

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,

 doi:10.20944/preprints202510.1615.v1

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.

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.

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.

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/.

Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016a. URL http://arxiv.org/abs/1604.08608.

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.

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.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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.

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*, 2024a.

A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.

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, 2020a.

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.

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.

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, 2020b.

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.

Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.

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.

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*, 2023a.

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.

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.

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, 2022a.

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.

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.

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, 2023b.

Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024a. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

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.

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.

Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024b. 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.

Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025a. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.

Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. 2025b. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379* (2025).

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.

Hao Fei, Yafeng Ren, and Donghong Ji. 2020a, A tree-based neural network model for biomedical event trigger detection, Information Sciences, 512, 175

Hao Fei, Yafeng Ren, and Donghong Ji. 2020b, Dispatched attention with multi-task learning for nested mention recognition, Information Sciences, 513, 241

Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2021, A span-graph neural model for overlapping entity relation extraction in biomedical texts, Bioinformatics, 37, 1581

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.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

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, 2021a.

D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

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, 2021b.

K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

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*, 2023a.

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*

 doi:10.20944/preprints202510.1615.v1

Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023c.

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.

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*, 2024b.

Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.

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, 2024c.

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.

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, 2022b.

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.

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.

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, 2023b.

S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.

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,*, 2024d.

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.

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.

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.

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, 2022c.

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.

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.

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, 2020c.

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.

 doi:10.20944/preprints202510.1615.v1

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, 2018b.

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, 2023d.

P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.

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, 2023c.

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, 2023a.

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, 2023e.

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, 2023d.

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.

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.

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