Knowledge-Enhanced Image Captioning with Adaptive Graph-based Multimodal Alignment and LLM

Guoyi Li[1,2], Die Hu[1,2], Haozhe Li[1,2], Zhongjiang Yao[1,2], Wei Mi[1,2], Zongzhen Liu[1,2], Xiaodan Zhang[1,2,*], Honglei Lyu[1,2],

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{liguoyi, hudie, yaozhongjiang, miwei, liuzongzhen, zhangxiaodan, lvhonglei}@iie.ac.cn

Abstract

Image captioning is crucial for multimodal understanding, bridging visual content and natural language. Despite recent advancements in Large Multimodal Models (LMMs), when faced with unseen entities or scenes in the open world, even when attempting to leverage learned knowledge, models still struggle with vague and inaccurate descriptions, and may even generate knowledge hallucinations. A key reason is that the model fails to effectively integrate knowledge with visual information, limiting its understanding of visual content. Thus, we propose Adaptive Knowledge Graph-guided Multimodal Alignment (AKGMA) for image captioning, which enhances semantic understanding in open-world scenes through visual knowledge reasoning, reducing knowledge hallucinations and improving caption quality. It consist three key components: Entity-guided Knowledge Aligner (EKA), Adaptive Knowledge Graph Construction (AKGC), and Scene-Context Knowledge Adapter (SCKA). EKA connects visual entities to knowledge graphs, providing structured knowledge to a small language model, which interacts with a visual encoder to acquire visual knowledge. AKGC uses reinforcement learning to build image-relevant subgraphs to optimize knowledge prompts and improve knowledge hallucinations. SCKA leverages scene graph annotations to extract visual contextual knowledge and inject it into Large Language Models (LLMs), ensuring the generated descriptions are consistent with the image's details. Additionally, we introduce UniKnowCap, a new image knowledge description dataset spanning various open-world knowledge domains, designed to evaluate the knowledge accuracy and detail consistency of model-generated descriptions. Extensive experiments show our model outperforms baselines across multiple metrics.

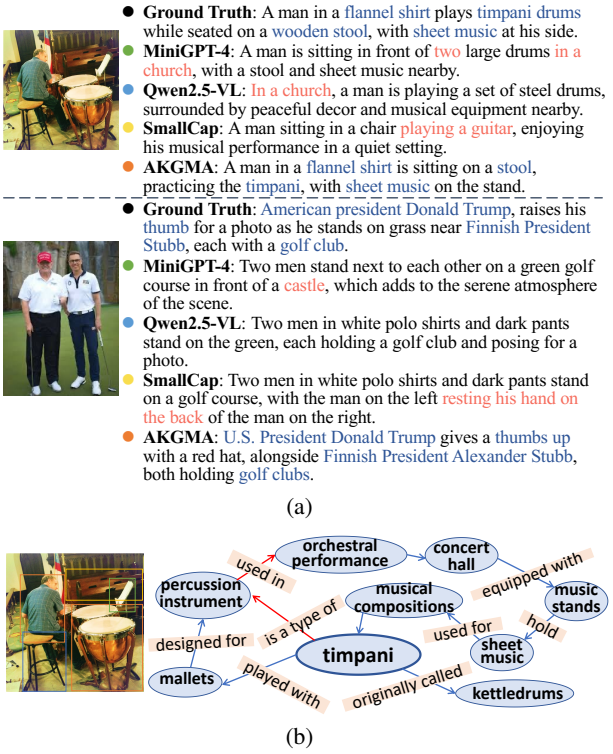

Figure 1: Examples illustrate our motivation.(a) Comparison of descriptions generated by various models, (b) Relevant knowledge graph. Incorrectly identified objects are in red, correct ones in blue.

Introduction

Image captioning, crucial for multimodal understanding, bridges visual content with natural language by generating semantically aligned descriptions. Recent methods leveraging Large Multimodal Models (LMMs) (Nichol et al. 2022; Saharia et al. 2022) have significantly improved performance (Li et al. 2023b). However, LMM-based approaches still face two critical limitations that hinder robust real-world deployment.

First, in the real world, many objects and scenes may not appear in the training data. Although the model tries to interpret the image using common knowledge concepts and scenes found in the pre-trained image-text dataset, this interpretation is often constrained by data quality issues (such as noise and inconsistencies in image-text pairs), leading to vague descriptions and even knowledge hallucinations. As shown in Figure 1 (a), both MiniGPT-4 and Qwen2.5-VL incorrectly identify the first example scene as a "church" and overlook certain specific knowledge concepts. While some methods attempt to incorporate external knowledge (Ramos

et al. 2023; Li et al. 2023b), models often mimic template-based knowledge text without truly improving open-world scene understanding—for instance, SmallCap may simply follow pre-existing templates and mistakenly assume the man in the image is "playing a guitar". Second, LMMs typically treat images as a "bag of objects", (Mitra et al. 2024), ignoring inter-object relations such as actions, spatial positioning, and interactions. This neglect leads to semantically inconsistent captions. As illustrated in Figure 1 (a, bottom), SmallCap incorrectly assumes that two people are riding piggyback.

Our motivation has two aspects: (1) Knowledge graphs are considered an effective tool to enhance understanding and reasoning capabilities through reasoning paths. As shown in Figure 1 (b), by recognizing 'timpani', a suitable reasoning path (red arrow) can reveal that it is commonly used in orchestra performances, rather than being misidentified as a "church" based on image style alone. Existing research shows that combining knowledge graphs with Large Language Models (LLMs) can enhance the knowledge reasoning ability of LLMs, thereby mitigating knowledge hallucinations (Sun et al. 2024; LUO et al. 2024). By selecting appropriate knowledge entities and relations, structured knowledge from knowledge graphs can be transformed into knowledge text prompts and injected into LLMs, improving the quality of generated image descriptions. (2) Scene graph (SG) annotations contain rich relationships and attributes between objects (Yang et al. 2022). We further enhance the consistency of generated captions by introducing a structured visual scene graph representation as contextual knowledge related to the image.

Thus, we propose Adaptive Knowledge Graph-guided Multimodal Alignment (AKGMA) for image captioning, consisting of Entity-guided Knowledge Aligner (EKA), Adaptive Knowledge Graph Construction (AKGC), and Scene-Context Knowledge Adapter (SCKA). To strengthen the model's visual semantic understanding, EKA integrates external knowledge with image representations through structured knowledge prompting and a small language model, projecting hidden states as knowledge embedding tokens that are injected into the LLM to enhance the model's deep understanding. Due to the potential vast size of knowledge bases, AKGC dynamically generates relevant knowledge subgraphs through reinforcement learning, working in conjunction with EKA to improve the quality of knowledge injected into the LLM. To ensure the alignment of descriptive details with visual information, SCKA extracts intra-image inter-object relational features (e.g., actions, spatial positions, and interactions), generating scene context vectors for each LLM layer. By integrating external and internal knowledge, AKGMA promotes semantic consistency between text and image, reducing knowledge hallucinations.

To evaluate the accuracy and generalization ability of captioning models, we introduce UniKnowCap, a large-scale dataset with 5,873 images and over 26,000 descriptions. Covering domains such as medicine, technology, and popular culture, many samples contain more than one knowledge entity and their interrelations, assessing the model's ability for cross-domain knowledge reasoning and complex scene description, providing a challenging benchmark.

Our contributions are: (1) We propose the Adaptive Knowledge Graph-guided Multimodal Alignment (AKGMA) for image captioning,which aligns visual features with structured knowledge (external knowledge and scene-context knowledge), extracts effective knowledge representations, and integrating it into LLMs to enhance knowledge recognition accuracy and semantic consistency in captions. (2) To select relevant knowledge entities and relations from extensive knowledge bases, we design the Adaptive Knowledge Graph Construction module to optimize structured knowledge prompts, improving visual knowledge quality and mitigating knowledge hallucinations. (3) We created UniKnowCap, a large-scale knowledge description dataset integrating multi-domain knowledge, to evaluate the knowledge accuracy and semantic consistency of image descriptions. Extensive experiments show our method outperforms baselines across various metrics.

Methodology

Our model consists of three main modules: (1) **Entity-guided Knowledge Aligner (EKA)**, which aligns visual features with knowledge generated from prompts and produces image-based knowledge representations in the LLM language space; (2) **Adaptive Knowledge Graph Construction (AKGC)**, which dynamically selects relevant knowledge subgraphs from external knowledge bases using reinforcement learning to refine structured prompts for EKA; and (3) **Scene-Context Knowledge Adapter (SCKA)**, which injects scene-specific knowledge into each LLM layer to ensure coherent and semantically enriched image descriptions. Specially, as shown in Figure 2, in the *Training Knowledge Graph Aligner with LLM*, EKA is trained to align visual features with relevant external knowledge graphs, thereby enhancing the integration of visual content and structured knowledge. In the *Generating Captions with Knowledge-Enhanced LLM*, adaptively selected external and internal context knowledge are injected into the LLM to produce high-quality, knowledge-enriched captions.

Entity-guided Knowledge Aligner

Entity Linkage. To improve captioning with external knowledge, we link image entities to corresponding ones in the knowledge graph (KG $\mathcal{G}$), converting image-based knowledge into textual prompts for the LLM. Mask R-CNN (He et al. 2017) detects objects in the image $\mathcal{I}$ (Qu, Tuytelaars, and Moens 2024), generating text queries to retrieve Wikipedia entities, which are then linked to the KG (Hu et al. 2023). The KG is defined as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ and $\mathcal{R}$ represent the entity and relation sets, and $\mathcal{T} = \{(e_s, r, e_t) \mid e_s, e_t \in \mathcal{E}, r \in \mathcal{R}\}$ is the set of triples.

Visual-Knowledge Alignment. In the training knowledge graph aligner with LLM phase, we design a visual knowledge generator that integrates structured knowledge with image features, producing visual knowledge representations mapped to the LLM's language space, forming the basis of EKA. Using CLIP for image representations $\mathbf{h}_{\mathcal{I}}$ and OPT-1.3B (Zhang et al. 2022) for visual-knowledge generation, similar to Q-former (Li et al. 2023a) and prefix tuning

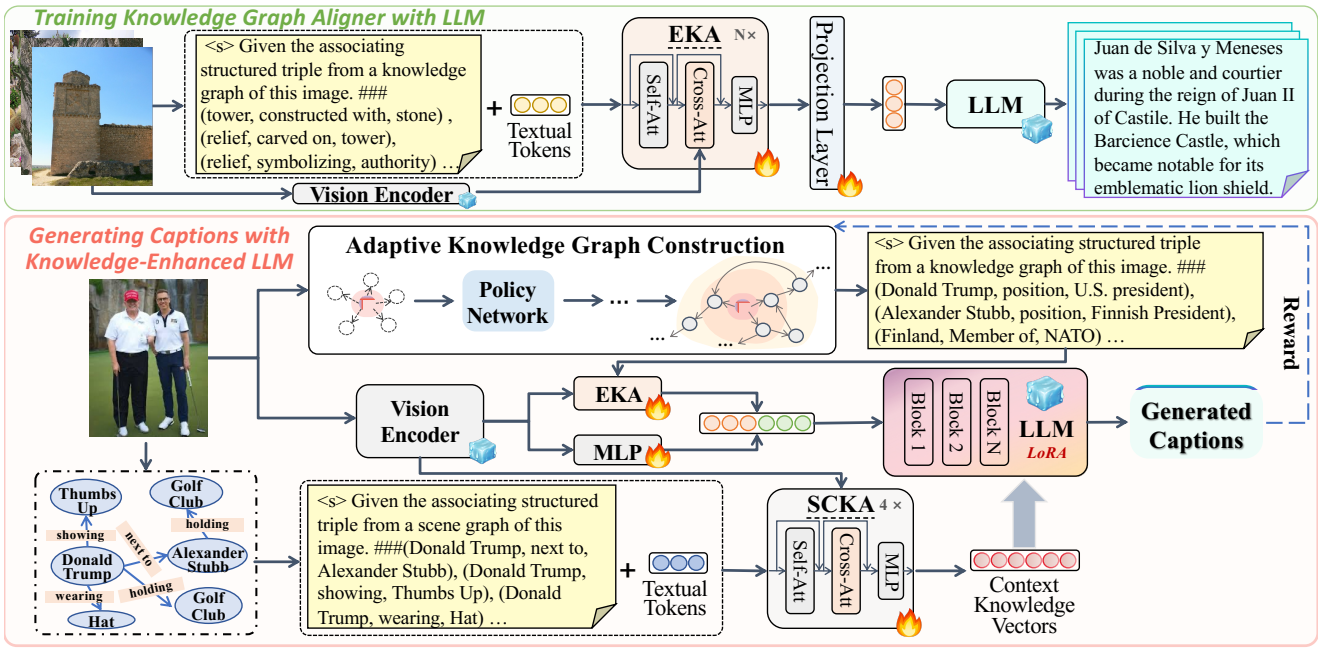

Figure 2: The overall architecture of AKGMA.

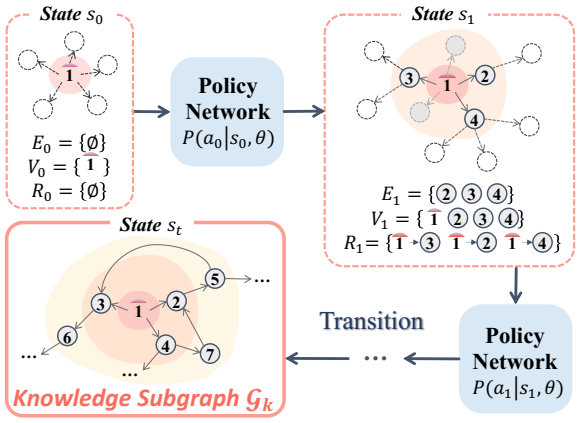

Figure 3: The MDP for reinforced adaptive generation of knowledge subgraphs.

(Li and Liang 2021), we align generated knowledge with visual content through an entity-driven approach. EKA is trained on the VKPairs dataset, built from the WIT dataset (Srinivasan et al. 2021), to align knowledge triples with visual knowledge, enhancing model reasoning by establishing structured paths between images and knowledge graph entities. We randomly select seed entities, expand them using random walk (Lu and Li 2012) to generate knowledge triples, and concatenate learnable token $\mathbf{h}_{kl}$ with knowledge prompts P_k: "<s> Given the associated structured triples from a knowledge graph of this image, describe the image in detail. ### (knowledge graph triples $\mathcal{T}_r$)". Cross-attention layers align visual representations of $\mathcal{I}$ with textual features, focusing on entities and relationships. A linear projection layer maps features into the language space:

$$\mathbf{h}_{vk}=\mathrm{EKA}([\mathbf{h}_{kp}||\mathbf{h}_{kl}],\ \mathbf{h}_{\mathcal{I}}),\quad \mathbf{h}_{ko}=\mathbf{W}_k\mathbf{h}_{vk}+\mathbf{b}, \tag{1}$$

where $\mathbf{h}_{vk}$ and $\mathbf{h}_{kp}$ are the visual knowledge generator output and knowledge prompt embeddings, respectively. $\mathbf{W}_k$ and $\mathbf{b}$ are learnable parameters connecting visual content with relevant knowledge from the graph and projecting it into the LLM. $\mathbf{h}_{ko}$ will be fed into the language models along with the original image representation $\mathbf{h}_{io}$. The latter is obtained through a learnable MLP trained on image-text captioning pairs, i.e., $\mathbf{h}_{io} = \mathrm{MLP}(\mathbf{h}_{\mathcal{I}})$. This module lays the foundation for adaptive knowledge graph-based alignment. By jointly training the two stages with distinct training corpora, our model can avoid early knowledge overfitting and achieves stronger generalization in open-domain scenarios. Joint training with the adaptive visual knowledge graph mapper prevents premature solidification of the model's knowledge, ensuring sufficient generalizability for diverse scenarios.

Adaptive Knowledge Graph Construction

EKA integrates structured knowledge with large models, but including all entities in prompts is impractical. We therefore propose adaptive knowledge subgraph generation, focusing on relevant entities to improve captions.

Given the complexity of enumerating all reasoning paths (Xu et al. 2020), we optimize the process with graph-based reinforcement learning (RL) (Park et al. 2022). The RL agent treats the knowledge graph as a framework, using visual context to guide reasoning and select relevant knowledge for EKA. Modeled as a Markov Decision Process (MDP) as shown in Figure 3, the agent constructs an optimized knowledge subgraph $\mathcal{G}_k$ by selecting actions that

maximize entity and relation relevance for caption generation.

The RL framework is modeled as an Markov Decision Process (MDP), where at each step the goal is to incorporate the most relevant knowledge for caption generation. The policy network processes actions, outputs probabilities, and drives state transitions. The agent explores knowledge entities until a size condition is met, forming the knowledge subgraph $\mathcal{G}_k$.

States. At time t, the state is $s_t = (\mathcal{I}, \mathcal{E}_t, \mathcal{V}_t, \mathcal{R}_t)$, where $\mathcal{I}$ is the image, and $\mathcal{E}_t$, $\mathcal{V}_t$, and $\mathcal{R}_t$ represent newly added entities, absorbed entities, and relationships, respectively. The initial state is $s_0 = (\mathcal{I}, \{e_i\}, \{\}, \{e_i\})$, where e_i is the seed node. The state evolves by adding relevant entities and relationships for context-aware captions.

Action. The action space $\mathcal{A}_t$ for state s_t consists of all outgoing relationships from entities in $\mathcal{E}_t$: $\mathcal{A}_t = \{(r, e) | (e_n, r, e) \in \mathcal{G}, e_n \in \mathcal{E}_t, e \notin \mathcal{V}_t\}$. The policy network predicts action probabilities, guiding transitions to the next state s_{t+1}. The agent selects entities and relationships relevant for caption generation, ensuring the subgraph supports precise captions while minimizing redundancy.

Transition. Given state $s_t = (\mathcal{I}, \mathcal{E}_t, \mathcal{V}_t, \mathcal{R}_t)$ and action $a_t = \{(r, e)\}$, the system transitions to state $s_{t+1} = (\mathcal{I}, \mathcal{E}_{t+1}, \mathcal{V}_{t+1}, \mathcal{R}_{t+1})$, where $\mathcal{E}_{t+1}$, $\mathcal{V}_{t+1}$, and $\mathcal{R}_{t+1}$ are updated based on the current state and action, refining the knowledge graph and focusing on relevant entities.

Subgraph construction and attention mechanisms are central to the function of the policy network $\pi_\theta(s_t, a_t)$, which estimates the conditional probability distribution of actions given state s_t. The state s_t is represented as: $s_t = (\mathbf{g}_t \| \mathbf{v}_t \| \mathbf{e}_i)$, where $\|$ denotes the concatenation operation. Here, $\mathbf{e}_i$, consistent with $\mathbf{h}_\mathcal{I}$ above, represents the image $\mathcal{I}$, while $\mathbf{g}_t$ is the vector representation of the subgraph at step t, and $\mathbf{v}_t$ denotes the vector representation of newly added entities at step t. The computation of $\mathbf{g}_t$ consists of refining the entity embeddings $\mathbf{e}_n$ within their neighborhood $\mathcal{V}_t$, followed by aggregating all entities using an attention network. The process begins with refining the entity embeddings:

$$\mathbf{e}'_n = \tanh(\mathbf{W}_0^\theta [\mathbf{e}_n \| \sum_{e_t \in \vartheta_t(e_n)} \alpha_{h,t} \mathbf{e}_t \| \mathbf{e}_i]), \tag{2}$$

where $\vartheta_t(e_n)$ denotes the neighbors of e_n. The attention weight $\alpha_{h,t}$ is computed to measure the relevance between entities e_h and e_t in the context of the image:

$$\alpha_{h,t} = \frac{\exp(\text{sim}(e_n, e_t, e_i))}{\sum_{e_t \in \mathcal{V}_t(e_n)} \exp(\text{sim}(e_n, e_t, e_i))}. \tag{3}$$

The similarity function $\text{sim}(e_n, e_t, e_i)$ incorporates the image representation e_i:

$$\text{sim}(e_n, e_t, e_i) = \text{ReLU}(\mathbf{e}_n^\top \mathbf{W}_v^\theta \mathbf{e}_i) + \text{ReLU}(\mathbf{e}_t^\top \mathbf{W}_v^\theta \mathbf{e}_i). \tag{4}$$

Using the refined entity embeddings $\mathbf{e}'_n$, the attention weights are recalculated: $\alpha_0^\theta(e_n) = \mathbf{w}^\alpha \text{ELU}(\mathbf{W}_1^\theta \mathbf{e}'_n)$, where ELU is the exponential linear unit. The normalized attention weights $\alpha^1(e'_n)$ are computed as:

$$\alpha_1^\theta(e'_n) = \frac{\exp(\alpha_0^\theta(e'_n))}{\sum_{e_c \in \mathcal{E}_t} \exp(\alpha_0^\theta(e_c))}. \tag{5}$$

Finally, the vector representation $\mathbf{g}_t$ is obtained by aggregating the refined entity embeddings $\mathbf{e}'_h$: $\mathbf{g}_t = \sum_{e_n \in E_t} \alpha_1^\theta(e'_n) \mathbf{e}'_n$. The embedding of newly added entities v_t is computed by averaging the entities within $\mathcal{E}_t$: $\mathbf{v}_t = \frac{\sum_{e_n \in \mathcal{E}_t} \mathbf{e}_n}{|\mathcal{E}_t|}$. Each available action a_i is represented as: $\mathbf{a}_i = (\mathbf{r}_i + \mathbf{e}_i)$, denoting the superposition of two vectors. To accelerate the generation process, we let the agent absorb multiple outgoing relationships in one action a_t, instead of absorbing only one relationship (Jiang et al. 2020). The policy network then calculates the probability of taking an action unit a_i at state s_t as follows:

$$\pi_\theta(s_t, a_i) = \frac{\exp(\mathbf{W}_2^\theta \text{ReLU}(\mathbf{W}_3^\theta [\mathbf{s}_t \| \mathbf{a}_i]))}{\sum_{a_j \in \mathcal{A}_t} \exp(\mathbf{W}_2^\theta \text{ReLU}(\mathbf{W}_3^\theta [\mathbf{s}_t \| \mathbf{a}_j]))}. \tag{6}$$

Reward. The reward function evaluates navigational states using Image Relevance and Terminal Rewards. Image Relevance guides entity selection, while Terminal Reward assesses caption accuracy.

(i) The agent selects relevant entities from the current image–caption pair $(\mathcal{I}, c_{\text{gt}})$, linking knowledge graph entities to VKPairs via an entity-to-sample mapping $\Lambda(e)$, where $\Lambda(e) = \{j \mid e$ appears in the knowledge background text of sample $j\}$. The reward is defined as the cosine similarity between the current multimodal representation and the averaged multimodal representations of the retrieved samples:

$$\mathcal{R}^B = \sum_{e_i \in a_m} \mathcal{R}'(n, e_i) = \cos\left(\mathbf{m}_n, \frac{\sum_{j \in \Lambda(e_i)} \mathbf{m}_j}{|\Lambda(e_i)|}\right), \tag{7}$$

where $\mathbf{m}_n = \text{CLIP}(\mathcal{I}, c_{\text{gt}})$ is the multimodal representation of the current image–caption pair, and $\mathbf{m}_j = \text{CLIP}(\mathcal{I}_j, c_{\text{gt},j})$ denotes the multimodal representations of VKPairs samples linked to entity e_i, prioritizing entities whose historical multimodal contexts are better aligned with the current ground-truth image semantics.

(ii) Terminal Reward: Applied at the final step T, after constructing the knowledge subgraph, it uses the subgraph as a prompt to improve semantic coherence in the generated caption and reduce irrelevant information. The Log Probability Reward (Dessì et al. 2023) measures the match between the image and generated description. For target image $\mathcal{I}$ and set $D \cup \{\mathcal{I}\}$, the reward is:

$$\mathcal{R}_s^T = \log \frac{\exp(\cos(E_t(c), E_i(\mathcal{I})))}{\sum_{\mathcal{I}' \in D \cup \{\mathcal{I}\}} \exp(\cos(E_t(c), E_i(\mathcal{I}')))}, \tag{8}$$

where $E_t(\cdot)$ and $E_i(\cdot)$ are CLIP text and visual encoder representations, and $D \cup \{\mathcal{I}\}$ includes the target image and 99 distractors. To encourage faithful and informative caption generation, we additionally incorporate a CIDEr-based reward $\mathcal{R}_c^T = \text{CIDEr}(c, c_{\text{gt}})$. The terminal reward is thus: $\mathcal{R}^T = \mathcal{R}_s^T + \mathcal{R}_c^T$, which jointly promotes semantic alignment and descriptive quality. Maximizing this reward improves alignment between the caption and image. The total reward $\mathcal{R}(t)$ is:

$$\mathcal{R}(t) = \lambda \cdot \mathcal{R}^B(t) + (1 - \lambda) \cdot \mathcal{R}^T(t), \tag{9}$$

where λ balances the relevance and terminal rewards.

Scene-Context Knowledge Adapter

To improve LLMs' understanding of image-object relationships for better contextual coherence, scene-based knowledge is essential (Mitra et al. 2024). We propose the

Scene-Context Knowledge Adapter (SCKA), which integrates scene knowledge into each LLM layer. The image is transformed into a scene graph $\mathcal{G}_c = \{\mathcal{E}_c, \mathcal{R}_c, \mathcal{T}_c\}$ using a pre-trained model (Tang et al. 2020), representing entities, relationships (e.g., "near", "holding"), and triples $\mathcal{T}_c$. Learnable tokens $\mathbf{h}_{sk}$ are combined with scene graph prompts P_s: "<s> Given the scene graph triples, describe the image. ### ($\mathcal{T}_s$)", integrating scene knowledge into the LLM. The SCKA computation is:

$$\mathbf{h}_{sk}^{l-1}=\text{Self-Att}(\mathbf{h}_{sc},\mathbf{h}_{cg})+\mathbf{h}_{SCKA}^{l-1}, \tag{10}$$

$$\mathbf{h}_{cr}=\text{Cross-Att}(\mathbf{h}_{sk}^{l-1},\ \mathbf{h}_{\mathcal{I}})+\mathbf{h}_{sk}^{l-1}, \tag{11}$$

$$\mathbf{h}_{SCKA}^{l}=\text{MLP}(\mathbf{h}_{cr}), \tag{12}$$

where $\mathbf{h}_{SCKA}^{l-1}$ is the previous SCKA output, and Self-Att and Cross-Att refer to self-attention and cross-attention operations. Refined visual knowledge $\mathbf{h}_{SCKA}$ is then injected into each LLM block for deeper interaction with context knowledge.

As shown in Figure 2 (bottom right), the first LLM layer takes $\mathcal{I}$, $\mathbf{h}_{ko}$, and $\mathbf{h}_q$ as inputs. The sequence $\mathbf{h}_{SCKA}$ is spliced across layers, producing vectors $\mathbf{h}_{SCKA}^1, \ldots, \mathbf{h}_{SCKA}^{LN}$, where LN is the total number of layers. These vectors are prepended to the input sequence at each layer, ensuring progressive integration of scene-context knowledge. This framework efficiently aligns visual knowledge and injects scene-based knowledge via EKA and SCKA.

The Optimization Framework

In this section, we introduce the multi-task optimization framework to train the AKGMA model.

Actor-Critic Optimization. The critic estimates the action value function $Q(s, a)$ in the MDP environment, with the network defined as:

$$Q_\phi(s_t,a_i)=\mathbf{W}_Q^0\text{ReLU}(\mathbf{W}_Q^1[\mathbf{s}_t\|\mathbf{a}_i]). \tag{13}$$

The critic is trained using Temporal Difference (TD) learning (Sutton 1988), with the target q_t calculated by the Bellman equation (Denardo 2012):

$$q_t=\mathcal{R}(t)+\mathbb{E}_{a\sim\pi_\theta}[\gamma\cdot Q_\phi(s_{t+1},a)], \tag{14}$$

where γ is the decay factor. The TD error is minimized via: $\mathcal{L}_{critic} = (Q_\phi(s_t, a_t) - q_t)^2$. The actor maximizes the expected reward, $J_{actor}(\theta) = \mathbb{E}_{a\sim\pi_\theta}[Q_\phi(s_t, a)]$, using the policy gradient method (Sutton et al. 1999), with the gradient:

$$\nabla_\theta J_{actor}(\theta)\simeq Q_\phi(s_t,a_t)\cdot\nabla_\theta\log\pi_\theta(s_t,a_t). \tag{15}$$

Cross-Entropy Loss. We train on the visual-knowledge pairs dataset VKPairs, aligning visual content with structured external knowledge, enhancing the model's reasoning by creating paths between image content and knowledge. Additionally, we train on the widely-used MSCOCO dataset for optimizing caption generation across real-world images. The Entity Knowledge Aligner (EKA) is trained on knowledge-rich visual-knowledge pairs with knowledge description token length $\mathcal{L}_k$. The cross-entropy loss $\mathcal{L}_{s_1}$ is:

$$\mathcal{L}_{s_1}=-\sum_{i=1}^{L_k}\log P_i(\hat{y}_i=y_i|\mathbf{h}_I,\mathbf{h}_{kp},\mathbf{h}_{kl};y_1,\ldots,y_{i-1}). \tag{16}$$

For MSCOCO, the standard cross-entropy loss $\mathcal{L}_{s_2}$ is:

$$\mathcal{L}_{s_2}=-\sum_{i=1}^{L_c}\log P_i(\hat{y}_i=y_i|\mathcal{I},\mathrm{P}_c;l_1,\ldots,l_{i-1}), \tag{17}$$

where L_c is the total token count and l_i is the i-th token.

The overall training objective is: $\mathcal{L}_{AKGMA} = \psi \cdot \mathcal{L}_{s_1} + \mathcal{L}_{s_2}$, where ψ is the balancing coefficient. Additional details can be found in the supplementary materials.

The UniKnowCap Dataset

To robustly evaluate knowledge-enhanced image captioning, we introduce UniKnowCap, a large-scale, multi-domain dataset. Unlike MSCOCO (Chen et al. 2015), Flickr (Plummer et al. 2015), and Nocaps (Agrawal et al. 2019) (focused on common objects) or KnowCap (Cheng et al. 2023) (limited in domain and complexity), UniKnowCap offers substantially greater scale and diversity, providing a more challenging benchmark.

UniKnowCap contains 5,873 images and over 26k descriptions across diverse domains such as medicine, technology, food, popular culture, history, and arts, nearly four times the size of KnowCap's 1,424 images and 4,100 descriptions. For each image, we collected five carefully crafted human-written reference captions (20∼30 words). This multi-domain expansion enables a more comprehensive assessment of the model's generalization and cross-domain reasoning. Some images feature a mix of historical events, famous figures, landmarks, technological devices, and academic concepts, requiring accurate entity recognition and relationship inference. Following (Cheng et al. 2023), we used ChatGPT to generate domain knowledge keywords, crawled over 30k images, and selected 5,873 via expert annotation for domain diversity and reduced bias. We employed GPT-4V to assess UniKnowCap on Knowledge Relevance, Semantic Consistency, and Detail Richness (1∼5), with UniKnowCap achieving top results in all metrics, validating its utility for knowledge-grounded captioning.

Experiments

Datasets. We train AKGMA on the MSCOCO Training set (Lin et al. 2014) and VKPairs, derived from Wit (Srinivasan et al. 2021). AKGMA is evaluated on MSCOCO Test, KnowCap Test (Cheng et al. 2023), NoCaps val (Agrawal et al. 2019), and Flickr30k (Plummer et al. 2015) Test sets and our new UniKnowCap. VKPairs comprises 2M Wikipedia images with descriptions and associated background text, processed via TFRecord parsing and thread-pool downloads, then filtered to remove 1% invalid URLs.

Baselines. Comparisons include LLM-based zero-shot methods: OFA (Wang et al. 2022), CapDec (Nukrai, Mokady, and Globerson 2022), MiniGPT4 (Zhu et al. 2024), Qwen2.5-VL (Bai et al. 2025); and fine-tuning methods: SmallCap (Ramos et al. 2023), ViECap (Fei et al. 2023), EVCap (Li et al. 2024), FuseCap (Rotstein et al. 2024);Qwen2.5-VL† (Bai et al. 2025), BLIP$_{\text{K-Replay}}$ (Cheng et al. 2023), OFA$_{\text{K-Replay}}$ (Cheng et al. 2023), ClipCap (Mokady, Hertz, and Bermano 2021).

Metrics and setup. We evaluate using BLEU (B@4), METEOR (M) , ROUGE (R), CIDEr (C), and SPICE (S).

Method	LLM	KnowCap Test							MSCOCO Test					NoCaps val (CIDEr)			
		B@4	M	R	C	Rec	$C_S \downarrow$	$C_I \downarrow$	B@4	M	C	S	CLIP-S↑	In	Near	Out	Overall
LLM-based Zero-shot																	
OFA	OFA	13.4	13.3	30.0	50.0	43.8%	9.2	5.0	38.2	30.8	130.7	23.8	0.82	-	-	-	-
CapDec	GPT-2	9.7	12.3	30.5	20.7	3.5%	7.7	4.5	26.8	25.2	92.5	11.6	0.71	33.2	36.2	48.7	38.3
Qwen2.5-VL	Qwen2.5-7B	17.3	15.6	38.2	58.1	41.7%	9.4	5.6	38.5	30.0	129.8	23.6	0.79	103.2	108.8	109.5	109.2
MiniGPT4	Vicuna-13B	15.7	14.8	32.5	50.1	38.3%	10.9	7.1	38.0	29.6	129.6	23.4	0.80	99.0	106.9	110.8	108.8
LLM-based Fine-tuning																	
ClipCap	GPT-2	21.5	18.3	37.3	66.5	35.2%	12.8	7.3	33.5	27.5	113.1	21.1	0.76	84.9	66.8	49.1	65.8
SmallCap	GPT-2	22.4	19.6	42.5	79.3	38.8%	11.6	7.0	36.3	28.2	121.3	21.7	0.78	87.9	84.6	84.4	85.0
FuseCap	BLIP	22.8	18.7	41.7	74.5	46.8%	11.3	6.9	36.8	28.6	124.3	22.1	0.79	93.9	91.6	90.4	95.3
ViECap	GPT-2	18.5	18.3	40.2	70.2	48.4%	8.9	5.3	26.5	24.2	93.5	18.2	-	65.3	71.3	73.2	70.3
EVCap	Viunca-13B	22.7	20.4	41.8	73.9	51.7%	9.4	6.0	41.5	31.2	140.1	24.7	-	111.7	119.5	116.5	119.3
Qwen2.5-VL†	Qwen2.5-7B	24.2	20.8	44.3	86.3	51.3%	9.8	6.0	40.4	30.5	136.4	24.4	0.84	113.8	112.6	112.3	113.4
$BLIP_{K\text{-}Replay}$	BLIP	22.3	19.6	43.3	81.8	50.3%	-	-	41.1	30.5	135.9	-	-	114.3	107.3	97.4	108.3
$OFA_{K\text{-}Replay}$	OFA	25.1	21.3	45.1	**99.6**	54.5%	9.7	5.9	40.1	31.6	138.1	24.6	0.87	-	–	–	–
AKGMA	Vicuna-7B	27.8	24.6	**46.7**	90.3	78.2%	7.1	4.0	42.8	**32.8**	146.6	**25.8**	0.90	**125.4**	120.5	124.6	120.8
AKGMA	Qwen2.5-7B	**28.5**	**25.5**	46.0	93.4	**81.5%**	**6.9**	**3.7**	**43.5**	31.6	**149.3**	25.5	**0.92**	124.8	**122.6**	**126.7**	**123.6**

Table 1: Comparison of methods on three benchmark image captioning datasets. Bold values indicate the best results, underlined values represent the best results among baselines, and "†" denotes using the same instruction tuning data for a fair comparison.

Method	UniKnowCap Test					Flickr30k Test	
	C	Rec	$C_S \downarrow$	$C_I \downarrow$	Natural↑	C	S
CapDec	28.7	7.9%	8.2	4.4	6.73	37.3	11.8
ViECap	56.3	37.2%	9.4	5.7	6.92	48.9	13.8
MiniGPT4	37.8	32.1%	11.2	7.3	7.56	77.2	16.9
$OFA_{K\text{-}Replay}$	76.7	42.3%	10.4	6.4	8.10	75.6	16.5
$BLIP_{K\text{-}Replay}$	72.6	46.8%	9.3	5.4	8.32	74.7	15.1
Qwen2.5-VL†	70.3	40.8%	10.2	6.2	7.84	76.4	16.2
w/o KG	76.0	48.0%	8.1	4.8	8.12	84.9	17.8
w/o (SCKA&AKGC)	75.2	54.8%	8.4	5.0	8.03	83.8	17.2
w/o AKGC	77.3	57.4%	8.0	4.5	8.27	85.8	18.2
w/o VKPairs	80.5	72.8%	7.8	4.6	9.02	86.3	18.0
w Static 1-order	80.7	54.2%	8.3	4.9	8.30	85.4	17.5
w Static 2-order	82.1	51.3%	8.9	5.3	8.42	84.7	17.0
AKGMA	**83.6**	**77.1%**	**7.3**	**4.0**	**9.26**	**88.3**	**18.8**

Table 2: Performance of different models on UniKnowCap and Flickr30k. "VKPairs" denotes the visual–knowledge pairs for training EKA. "Static 1-order/2-order" add 1-hop/2-hop knowledge neighbors. "w/o AKGC" replaces AKGC with random walk.

Knowledge Recognition Accuracy (RecogAcc) (Cheng et al. 2023) measures correct knowledge keywords, and CLIP-Score (CLIP-S) (Hessel et al. 2021) measures image–text semantic alignment. CHAIR (Rohrbach et al. 2018) evaluates object hallucination at sentence ($CHAIR_S$) and image ($CHAIR_I$) levels. The Natural metric uses GPT-4V (1–10) (Achiam et al. 2023) on UniKnowCap (Zhao et al. 2023) for 500 sampled images, focusing on object, color, location, and relation errors. We adopt CLIP (Radford et al. 2021) as the pre-trained text–image encoder and a scene graph generation tool (Tang et al. 2020) with a Mask-RCNN backbone for entity relations. Main experiments report two AKGMA variants (Vicuna-7B (Chiang et al. 2023) and Qwen2.5-7B (Bai et al. 2025)); other experiments use Qwen2.5-7B (LD=4),

Method	MSCOCO Test					Flickr30k Test	
	B@4	M	C	S	CLIP-S↑	C	S
w/o KG	41.1	29.5	138.2	24.4	0.87	87.3	18.2
w/o AKGC	41.9	30.1	142.2	24.8	0.89	86.0	18.0
w/o VKPairs	42.7	31.0	143.7	25.3	0.92	87.0	18.6
w/o SCKA	42.4	30.4	145.1	25.0	0.85	86.4	17.7
AKGMA (LD=2)	42.7	30.7	144.7	**25.8**	0.90	86.7	**19.1**
AKGMA (LD=4)	**43.5**	**31.6**	**149.3**	25.5	**0.92**	**88.3**	18.8
AKGMA (LD=8)	43.2	31.2	148.5	25.6	0.91	87.2	18.6

Table 3: Performance of different variants on MSCOCO and Flickr30k. "LD" denotes the length of context knowledge vectors used in SCKG. "w/o SCKA" is the same as "LD=0".

where LD is the length of context knowledge vectors.

Results

Overall Performance. Tables 1 and 2 compare AKGMA with baselines, showing superior performance on both in-domain (MSCOCO Test) and out-of-domain (other datasets) settings. Table 1 shows that, compared to LLM-based zero-shot models, LLM-based fine-tuning models perform better across various metrics. Beyond recognition accuracy, fine-tuned OFA and Qwen2.5-VL improve on MSCOCO Test in CLIP-S, CIDEr, and other metrics, enhancing semantic consistency and language quality. The success of our approach may be attributed to: (1) synergy between adaptive knowledge graph construction and entity-guided aligner modules, which, via reinforcement learning, optimize entity selection and alignment, possibly boosting knowledge accuracy in complex scenarios; (2) strong generalization enabled by entity linking, avoiding dataset-specific bias and improving adaptation to novel scenes; (3) integrating scene graph visual/semantic relationships into each LLM layer, aligning captions with both visual content and deeper context, thus possibly reducing inconsistency and vagueness and improv-

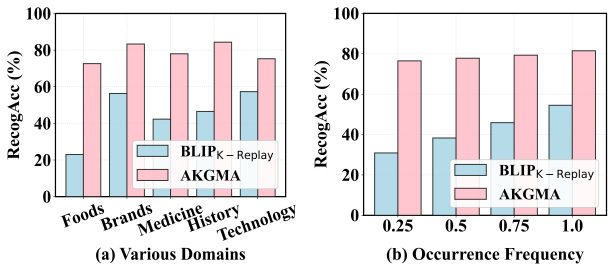

Figure 4: RecogAcc across scenario domains and its relation to occurrence frequency in pre-training data.

ing CIDEr/CLIP-S.

Ablation Study. Based on Tables 1, 2, and 3, we can conclude the following: (1) *Effects of Entity-guided Knowledge Aligner (EKA)*: Fine-tuning Qwen2.5-VL† under the same training conditions as AKGMA shows that AKGMA with EKA consistently outperforms in knowledge recognition accuracy, hallucination rate, and caption quality, compared to AKGMA without SCKA & AKGC. (2) *Effects of Adaptive Knowledge Graph Construction (AKGC)*: AKGC enhances caption relevance and knowledge accuracy by selecting contextually relevant entities. Replacing AKGC with static methods ("Static 1-order" and "Static 2-order") causes significant performance drops, increasing C_S and C_I due to noise and knowledge hallucinations. (3) *Effects of Scene-Context Knowledge Adapter (SCKA)*: SCKA improves CIDEr and SPICE scores by capturing scene context, but increasing context length (LD) excessively causes performance fluctuations, likely due to its negative impact on model structure.

Generalization Ability. To assess AKGMA's generalization across diverse domains, we evaluated its performance in UniKnowCap covering areas of brands, foods, technology, history and medicine. Figure 4 (a) shows AKGMA outperforming $\text{BLIP}_{\text{K-Replay}}$ in domains such as signature foods, medicine, and history. We hypothesize that AKGMA's use of relational associations within external knowledge graphs helps it recognize domain-specific features, unlike K-Replay, which depends on pre-training data. Analyzing knowledge frequency in pre-training data (Figure 4 (b)), $\text{BLIP}_{\text{K-Replay}}$ excels in frequently encountered domains but struggles with unseen entities, while AKGMA remains less influenced by data frequency, showing stronger stability.

Case Study

A UniKnowCap example (Figure 5) illustrates how AKGMA generates accurate, visually grounded captions. AKGMA adaptively constructs knowledge subgraphs and scene graph triples, injects image-relevant knowledge via EKA, and integrates contextual details from scene graphs (e.g.,"rides horse"). By filtering out irrelevant (gray) information and correctly linking entities (e.g., "Napoleon Bonaparte", "Berlin", and the "Brandenburg Gate"), AKGMA accurately captures the corresponding historical event and its timeline, and does not misidentify the building in the image as the "Arc de Triomphe" in Paris, thereby producing precise, contextually grounded captions.

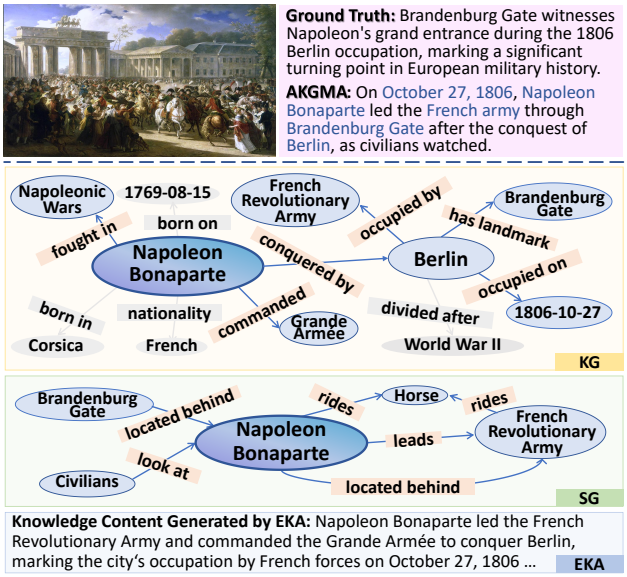

Figure 5: An example showcasing descriptions generated by our model.

Related Work

Existing works in image captioning focus on model improvements and leveraging external knowledge. **Non-LLM models** lack external knowledge and struggle with open-world generalization (Shi et al. 2020; Luo et al. 2021). **LLM-based models** leverage implicit knowledge but tend to generate generic captions for novel scenes (Li et al. 2022; Liu et al. 2024; Qu, Tuytelaars, and Moens 2024). **Knowledge-augmented approaches** use external or retrieved knowledge to enrich captions, but may introduce hallucinations and remain limited in scale or diversity (Rotstein et al. 2024; Cheng et al. 2023). **Knowledge graph-enhanced methods** inject structured knowledge for better reasoning, though may reduce interpretability; transforming KGs to text for LLM input is promising (Zhang et al. 2024; Dai et al. 2025).

Conclusion

We introduced AKGMA, an Adaptive Knowledge Graph-Guided Multimodal Alignment method that enhances knowledge reasoning, reduces hallucinations, and improves semantic consistency through visual context, representing an innovation over traditional visual-text alignment captioning models. We proposed the UniKnowCap dataset for evaluating model generalization in complex, knowledge-driven scenarios. Future work will explore knowledge injection strategies for caption diversity and semantic consistency.

Acknowledgments

This project is supported by the National Key Research and Development Program of China (No. 2022YFB2702500).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8948–8957.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Cheng, K.; Song, W.; Ma, Z.; Zhu, W.; Zhu, Z.; and Zhang, J. 2023. Beyond generic: Enhancing image captioning with real-world knowledge using vision-language pre-training model. In *Proceedings of the 31st ACM International Conference on Multimedia*, 5038–5047.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3): 6.

Dai, X.; Hua, Y.; Wu, T.; Sheng, Y.; Ji, Q.; and Qi, G. 2025. Large language models can better understand knowledge graphs than we thought. *Knowledge-Based Systems*, 113060.

Denardo, E. V. 2012. *Dynamic programming: models and applications*. Courier Corporation.

Dessì, R.; Bevilacqua, M.; Gualdoni, E.; Rakotonirina, N. C.; Franzon, F.; and Baroni, M. 2023. Cross-domain image captioning with discriminative finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6935–6944.

Fei, J.; Wang, T.; Zhang, J.; He, Z.; Wang, C.; and Zheng, F. 2023. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3136–3146.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7514–7528.

Hu, H.; Luan, Y.; Chen, Y.; Khandelwal, U.; Joshi, M.; Lee, K.; Toutanova, K.; and Chang, M.-W. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12065–12075.

Jiang, J.; Dun, C.; Huang, T.; and Lu, Z. 2020. Graph Convolutional Reinforcement Learning. In *International Conference on Learning Representations*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Li, J.; Vo, D. M.; Sugimoto, A.; and Nakayama, H. 2024. EVCap: Retrieval-Augmented Image Captioning with External Visual-Name Memory for Open-World Comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13733–13742.

Li, W.; Zhu, L.; Wen, L.; and Yang, Y. 2023b. DeCap: Decoding CLIP Latents for Zero-Shot Captioning via Text-Only Training. In *The Eleventh International Conference on Learning Representations*.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. 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)*, 4582–4597.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Lu, J.; and Li, D. 2012. Sampling online social networks by random walk. In *Proceedings of the First ACM international workshop on hot topics on interdisciplinary social networks research*, 33–40.

LUO, L.; Li, Y.-F.; Haf, R.; and Pan, S. 2024. Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning. In *The Twelfth International Conference on Learning Representations*.

Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.-W.; and Ji, R. 2021. Dual-level collaborative transformer for image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 2286–2293.

Mitra, C.; Huang, B.; Darrell, T.; and Herzig, R. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14420–14431.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022.

GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.

Nukrai, D.; Mokady, R.; and Globerson, A. 2022. Text-Only Training for Image Captioning using Noise-Injected CLIP. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4055–4063.

Park, S.-J.; Chae, D.-K.; Bae, H.-K.; Park, S.; and Kim, S.-W. 2022. Reinforcement learning over sentiment-augmented knowledge graphs towards accurate and explainable recommendation. In *Proceedings of the fifteenth ACM international conference on web search and data mining*, 784–793.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, 2641–2649.

Qu, T.; Tuytelaars, T.; and Moens, M. F. 2024. Visually-Aware Context Modeling for News Image Captioning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2927–2943.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramos, R.; Martins, B.; Elliott, D.; and Kementchedjhieva, Y. 2023. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2840–2849.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object Hallucination in Image Captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4035–4045.

Rotstein, N.; Bensaid, D.; Brody, S.; Ganz, R.; and Kimmel, R. 2024. Fusecap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 5689–5700.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494.

Shi, Z.; Zhou, X.; Qiu, X.; and Zhu, X. 2020. Improving Image Captioning with Better Use of Caption. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7454–7464.

Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; and Najork, M. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2443–2449.

Sun, J.; Xu, C.; Tang, L.; Wang, S.; Lin, C.; Gong, Y.; Ni, L.; Shum, H.-Y.; and Guo, J. 2024. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model on Knowledge Graph. In *The Twelfth International Conference on Learning Representations*.

Sutton, R. S. 1988. Learning to predict by the methods of temporal differences. *Machine learning*, 3: 9–44.

Sutton, R. S.; McAllester, D.; Singh, S.; and Mansour, Y. 1999. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12.

Tang, K.; Niu, Y.; Huang, J.; Shi, J.; and Zhang, H. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3716–3725.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, 23318–23340. PMLR.

Xu, K.; Song, L.; Feng, Y.; Song, Y.; and Yu, D. 2020. Coordinated reasoning for cross-lingual knowledge graph alignment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 9354–9361.

Yang, J.; Ang, Y. Z.; Guo, Z.; Zhou, K.; Zhang, W.; and Liu, Z. 2022. Panoptic Scene Graph Generation. In *European Conference on Computer Vision*, 178–196.

Zhang, S.; Roller, S.; Goyal, N.; Artetxe, M.; Chen, M.; Chen, S.; Dewan, C.; Diab, M.; Li, X.; Lin, X. V.; Mihaylov, T.; Ott, M.; Shleifer, S.; Shuster, K.; Simig, D.; Koura, P. S.; Sridhar, A.; Wang, T.; and Zettlemoyer, L. 2022. OPT: Open Pre-trained Transformer Language Models. arXiv:2205.01068.

Zhang, Y.; Chen, Z.; Guo, L.; Xu, Y.; Zhang, W.; and Chen, H. 2024. Making large language models perform better in knowledge graph completion. In *Proceedings of the 32nd ACM international conference on multimedia*, 233–242.

Zhao, Z.; Wang, B.; Ouyang, L.; Dong, X.; Wang, J.; and He, C. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *ICLR*.