

Controllable Contextualized Image Captioning: Directing the Visual Narrative through User-Defined Highlights

Shunqi Mao¹, Chaoyi Zhang¹, Hang Su², Hwanjun Song³, Igor Shalyminov², and Weidong Cai¹

¹ The University of Sydney

² AWS AI Labs

³ Korea Advanced Institute of Science and Technology

Abstract. Contextualized Image Captioning (CIC) evolves traditional image captioning into a more complex domain, necessitating the ability for multimodal reasoning. It aims to generate image captions given specific contextual information. This paper further introduces a novel domain of Controllable Contextualized Image Captioning (Ctrl-CIC). Unlike CIC, which solely relies on broad context, Ctrl-CIC accentuates a user-defined highlight, compelling the model to tailor captions that resonate with the highlighted aspects of the context. We present two approaches, Prompting-based Controller (\mathcal{P} -Ctrl) and Recalibration-based Controller (\mathcal{R} -Ctrl), to generate focused captions. \mathcal{P} -Ctrl conditions the model generation on highlight by prepending captions with highlight-driven prefixes, whereas \mathcal{R} -Ctrl tunes the model to selectively recalibrate the encoder embeddings for highlighted tokens. Additionally, we design a GPT-4V empowered evaluator to assess the quality of the controlled captions alongside standard assessment methods. Extensive experimental results demonstrate the efficient and effective controllability of our method, charting a new direction in achieving user-adaptive image captioning. Code is available at <https://github.com/ShunqiM/Ctrl-CIC>.

Keywords: Contextualized image captioning · Large multimodal model · Controllable text generation

1 Introduction

Image captioning stands as a crucial task that bridges computer vision (CV) and natural language processing (NLP) domains, aiming to create sentences that effectively summarize visual content. Benefiting from the supervised training on extensively annotated datasets [23, 37, 45], various learning-based methods [9, 18, 34] have been developed to leverage diverse neural architectures for more accurate image captions. The recent emergence of large language models (LLMs) [1, 8, 50] marks an arising paradigm in the field, excelling in a variety of complex multimodal tasks under both zero-shot [13, 27, 49] and few-shot [19, 52] scenarios, with image captioning as their fundamental component. This trend

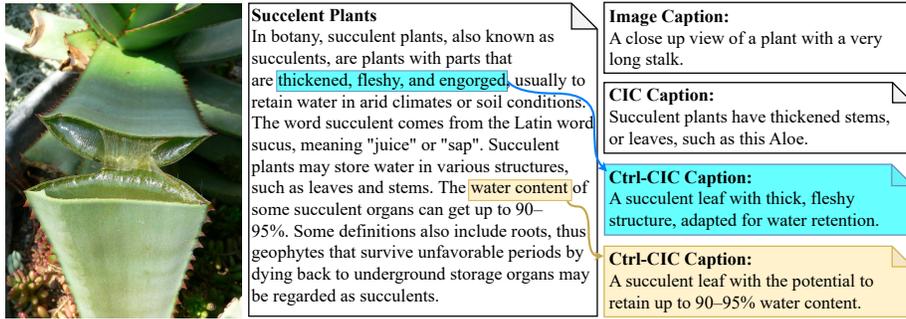


Fig. 1: We introduce the Controllable Contextualized Image Captioning (Ctrl-CIC) task: Given a global context, Ctrl-CIC aims at generating contextualized image captions tailored to specific highlighted segments. In the presented context regarding "succulents", highlights direct the caption’s emphasis, underscoring distinct attributes such as its anatomical structure or water content.

highlights a growing demand for more challenging image captioning tasks to be established, which could act as evolving test-beds for benchmarking the multi-modal reasoning capabilities of these cutting-edge LLM systems.

However, images are replete with rich visual cues encompassing diverse objects and varying levels of information granularity [15]. Specifically, a single image can be interpreted and captioned in numerous ways depending on the focused visual aspects within a given context, resulting in a low inter-annotator agreement. Consequently, generating captions for an image without considering specific context is often less controlled and ill-suited for real-world applications. Recognizing this, Nguyen *et al.* [31] recently introduced the Wikipedia captioning (WikiCap) task to assist Wikipedia authors in captioning images based on the contextual backdrop of the Wikipedia page in which they appear. Later, Burns *et al.* [5] expanded and formulated this WikiCap task into a more formal and generalized notion as Contextualized Image Captioning (CIC), wherein an image is captioned along with its relevant context.

CIC has significantly enhanced the utility of image captioning models in real-world applications. However, the complexity of context presents a notable challenge: a detailed context may not yield a single “ideal” caption, but rather a spectrum of valid captions that resonate with different elements within the image. In contexts like Wikipedia articles, where topics often have multiple facets, an image might align with multiple captions, each reflecting a unique aspect spotlighted in the associated text. Nevertheless, a mere truncation of the context to pinpoint intended highlights is not a viable solution. Such an approach, as shown in our experiments, would inadvertently omit crucial background or cues essential for the image in the context, thereby impeding the ability of the model to generate a comprehensive and precise caption.

To tackle this, we propose the Controllable Contextual Image Captioning (Ctrl-CIC) task. While CIC models often deal with diverse contexts, leading to multiple plausible captions, Ctrl-CIC introduces a user-controlled captioning

mechanism, enabling users to pinpoint specific highlighted segments of the context that the model should prioritize, as illustrated in Fig. 1. By integrating the advantages of standard CIC with this novel controllability, Ctrl-CIC harnesses the robust capabilities of deep models while leveraging the specificity of human intent, forging a path toward more refined and intentional captioning.

Furthermore, we present two preliminary methods for Ctrl-CIC: Prompting-based Controller ($\mathcal{P}\text{-Ctrl}$) and Recalibration-based Controller ($\mathcal{R}\text{-Ctrl}$). These simple yet effective approaches steer models to produce captions that are relevant to specified highlights. The $\mathcal{P}\text{-Ctrl}$ method utilizes a highlight-based prefix as a prompt to guide controllable caption generation, while $\mathcal{R}\text{-Ctrl}$ recalibrates the decoder weightings so that the generated contents are augmented towards highlights. Applicable to any LMs, these methods enable controlled caption generation with finetuning. To mitigate the lack of fine-grained highlights for Ctrl-CIC task, we further propose a weakly-supervised solution that generates pseudo-highlights from the context to facilitate effective Ctrl-CIC model training.

Noticeably, evaluating Ctrl-CIC presents unique challenges, as it requires assessment beyond reference-based metrics such as ROUGE [26] and BLEU [36], due to the absence of specific highlights and benchmarks. To this end, we leverage GPT-4 [33] to create test samples by extracting diverse highlights from contexts, and we adopt reference-free metrics to evaluate the alignment, relevance, and diversity of the Ctrl-CIC captions. Moreover, we propose a GPT-4V empowered evaluator, leveraging chain-of-thought techniques for comprehensive assessment of caption quality across multiple dimensions. Empirical results indicate that our approach is capable of generating captions that are not only diverse but also exhibit significant controllability in response to different contextual highlights.

To summarize, our key contributions are four-fold: (1) We formally introduce the Controllable Contextualized Image Captioning (Ctrl-CIC) problem, emphasizing the challenges of contextually influenced caption generation. (2) We present novel technical solutions, Prompting-based Controller ($\mathcal{P}\text{-Ctrl}$) and Recalibration-based Controller ($\mathcal{R}\text{-Ctrl}$), designed for the Ctrl-CIC task which aligns contextual captions with user intents. (3) We propose an extensive evaluation pipeline, employing GPT-4(V) for highlight selection and as an additional evaluator, alongside a set of subjective measures, to comprehensively assess the controllability and overall performance of the Ctrl-CIC models. Empirical analysis demonstrates the effectiveness of the proposed controllers in Ctrl-CIC task, outperforming large vision-language models with much fewer parameters.

2 Related Work

Contextualized Image Captioning. Traditional image captioning methods typically employ an encoder-decoder architecture to transform visuals into texts [2, 32, 35]. Contextualized Image Captioning (CIC) [31] incorporates a textual context to the captioning task, aiming at generating captions that are not only descriptive but also contextually relevant. The CIC task can be traced back to news image captioning [42], a similar task that aims to generate descriptions for

images embedded in news articles. Biten *et al.* [4] proposed a founding method of news image captioning by extracting named entities from context and filling them into generated caption templates. Following works continue to explore recognizing named entities [39,51,59,62] or extracting factual knowledge from the context [57,61,65], as well as large-scale pretraining [20,31], to facilitate context-based captioning. The most relevant work to ours, [5], proposed an efficient prefix attention mechanism to handle the lengthy context, together with a large-scale CIC dataset that extends CIC to a more generalized domain instead of news articles. However, the context often contains redundant text or excessive information that can be equally relevant to the image. In Ctrl-CIC, we guide the model with context highlights to ensure it generates highlight-specific captions.

Vision-Language Model for Image Understanding. The latest state-of-the-art approaches on image captioning are dominated by finetuned large Vision-Language Models (VLMs) that bridge the gap between vision and language [24, 25, 53, 56]. Similarly, the task of image captioning under contextual or controllable setup can be aptly addressed using VLMs. The majority of these models rely heavily on a vast corpus of image-text pairs for training. For example, GIT [53] leverages a generative image-to-text transformer trained on 0.8B image-text pairs. Conversely, Blip2 [25] maintains static image and language models, focusing instead on training a lightweight connection module, Qformer, using 129M images. Additionally, there is a subset of VLMs like Flamingo [1] and FRO-MAGe [22] that interleave image-text sequences in a unified style. Beyond mere training, there is a growing interest in creating general-purpose vision-language models finetuned with instructions. InstructBlip [10] augments the capabilities of Blip2 by introducing instruction-aware visual feature extraction attention. Similarly, models like LLava [27] and MiniGPT-4 [66] are designed for interactive conversations with multimodal data. However, while these models are often heavy in terms of computational resources, our approach stands out for its efficiency, requiring only 5% of the parameters compared to VLMs such as LLaVA, while providing enhanced performance.

Controllable Text Generation. Differing from standard text generation that might produce outputs based purely on training data or the input prompt, controllable text generation (CTG) allows for deriving more directed and purposeful outputs, by flexibly adjusting user-defined conditions. A line of CTG approaches emphasizes modulating the semantics of the produced content, focusing on attributes such as sentiment [7, 44, 48, 54, 60], topic [11, 21, 58], and persona [46, 47, 64]. Further research delves into lexical control during generation, for instance, integrating specific keywords or phrases [6, 16, 63]. Length-controlled language models have also been a subject of considerable investigation [12, 29, 30]. Moreover, structured control elements, like tables and trees, have also been a topic of interest in research [38, 43]. Traditional methods often rely on fixed categories, templates, or simple keyword rules, limiting flexibility. Our approach allows for using any contextual highlights as guidance, improving adaptability.

While alignment with these highlights is crucial, we do not enforce their direct inclusion in the output.

3 Method

3.1 Revisiting Contextualized Image Captioning

Extending traditional image captioning into a domain that demands multimodal reasoning capabilities, CIC offers a more practical setup by incorporating additional contextual information into captioning. This context-driven method could largely benefit many real-world cases where images are often accompanied by surrounding textual content, such as news [42] or Wikipedia captioning [31].

Formally, for a given image I and its context C , where C is a set of language tokens represented as $C = \{c_1, c_2, \dots, c_n\}$, Contextual Image Captioning aims at generating captions \mathcal{G} as:

$$\mathcal{G} = M_{\text{CIC}}(I, C), \quad (1)$$

where $M_{\text{CIC}}(\cdot, \cdot)$ denotes the vision-language model leveraging both the image and text. A CIC dataset typically contains triplet-formed data samples D_{CIC} , which can be denoted as:

$$D_{\text{CIC}} = \{(C, I, T), \dots\}, \quad (2)$$

where each triplet consists of an image I , its corresponding context C , and the target caption T . Such datasets allow models to learn the suitable caption for images under respective contexts.

However, one intuitive challenge of CIC is that context can sometimes be overflowing with redundant text or possess excessive information, adding a layer of complexity to distinguishing the most relevant details. Taking Fig. 1 as an example, given an image of a succulent paired with an article detailing the various characteristics of the plant, a CIC model may potentially generate a group of acceptable captions such as “A close-up view of a succulent’s stem” or “succulents leaves in detail”, each focusing on different aspects of the context.

To this end, we propose Ctrl-CIC and introduce the concept of “highlights”, directing the model to concentrate on specific aspects of the context. Ctrl-CIC ensures that the generated captions are more closely aligned with the human intent highlighted in these sections. Meanwhile, the rest of the context serves primarily to furnish the model with background knowledge. Consequently, the model is enabled to not only generate captions that are more pertinent to the highlighted segment but also to effectively utilize the broader context’s knowledge for enhanced relevance and accuracy.

3.2 Controllable Contextual Image Captioning

We first introduce the “highlights” notation H , which reflects a particular intent or focus within the context. The Ctrl-CIC objective now becomes:

$$\mathcal{G} = M_{\text{CCIC}}(I, C, H). \quad (3)$$

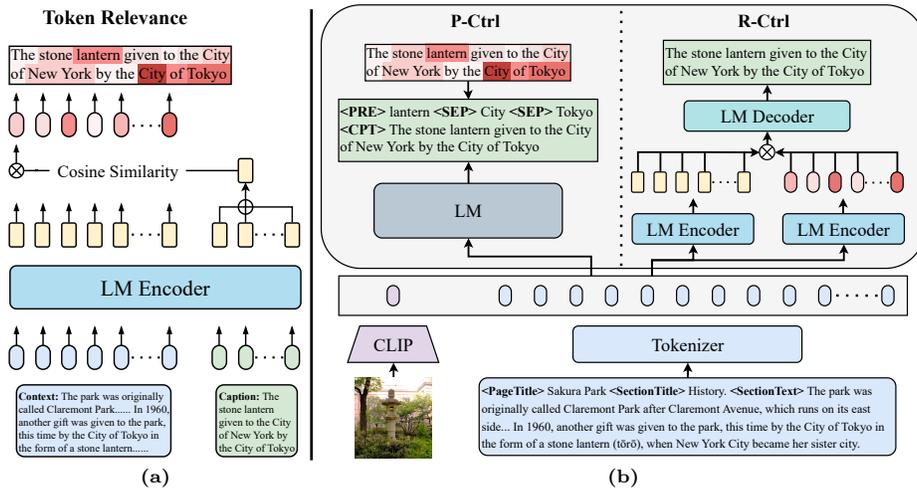


Fig. 2: Overview of the proposed Ctrl-CIC method. (a) We derive the token-level relevance scores that indicate the probability of the token being part of the highlights for the context-caption pair. (b) Overview of the training pipeline of the Prompting-based and Recalibration-based Controllers. For Ctrl-CIC inference, the model is guided by either new prompts or recalibrated weights based on highlights, to produce controlled captions.

Here, the generated caption \mathcal{G} is particularly influenced by the highlighted context H , while still taking into account the overall context C . Ideally, such Ctrl-CIC model should be trained on a dataset of the form:

$$D_{\text{CIC}} = \{(C_i, I_i, H_i^j, T_i^j), \dots\}, \quad (4)$$

where i denotes the index of each context-image pair in the dataset, and the other index j specifies the j^{th} associated highlight H_i^j and the target caption T_i^j of the i^{th} data sample. Such H should satisfy the following criteria:

1. The highlights, $H = \{h_i\}$, is a collection with each member h_i being a continuous subsequence of C .
2. Each highlighted component h_i within H could range from individual words to complete sentences, emphasizing our requirement for h_i to possess granularity at least down to the word level.

While the target caption T for traditional CIC is provided naturally in many paired image-text datasets and can be easily collected for training a standard CIC model, annotating sufficient amount of paired highlights and captions (H , T) for Ctrl-CIC can be manpower-intensive, due to the scarcity of diverse captions corresponding to a single image and context on the Internet. To the best of our knowledge, we could not identify any dataset with annotations that could fit or be adapted to this description. Given this limitation, we construct the control-

lable highlights H from a CIC dataset comprised of paired data (C, I, T) to train Ctrl-CIC models in a weakly supervised manner with the following method.

Intuitively, for a given context C , the highlight H controlling the generation of a Ctrl-CIC caption T should comprise highlighted components h that bear the highest relevance to the caption among other elements in C . Therefore, we aim to find such h , as illustrated in Fig. 2a.

Firstly, the context embedding $\text{Emb}_{ctx} \in R^{N_{ctx} \times 768}$ is derived through an encoder $\text{Enc}(\cdot)$ as $\text{Emb}_{ctx} = \text{Enc}(C)$, where N_{ctx} denotes the number of tokens in context C . Similarly, the embedding of the target caption $\text{Emb}_{tgt} \in R^{N_{tgt} \times 768}$ can also be encoded as $\text{Emb}_{tgt} = \text{Enc}(T)$, followed by an average pooling among tokens to derive its global embedding $\overline{\text{Emb}}_{tgt}$ and N_{tgt} denotes the number of tokens in the target caption. Next, we obtain the token-level relevance scores $S = \{s_i\}$ via computing the cosine similarity between the context and target caption as:

$$S = \{s_i = \cos(\text{Emb}_{ctx}^i, \overline{\text{Emb}}_{tgt}) \mid 0 \leq i < N_{ctx}\}. \quad (5)$$

The resulting token-level similarity score, S , implies the relevance between each context token and the caption. They can then serve as the candidate scores for each token in the context when deciding their inclusion in the highlight. Through aggregation, these token-level relevance scores can extend seamlessly to word-level granularity with averaging: $s_{w_j} = \frac{1}{k} \sum_{l=i}^{i+k-1} s_l$, where w_j denotes a single word spanning through k tokens (from c_i to c_{i+k-1}), and s_{w_j} indicates its word-level relevance score, thereby ensuring adaptability for highlight components h of diverse forms.

The following methods are designed to ensure that the content generation process is effectively conditioned on the highlights.

3.3 Controlled Caption Generation

We now introduce our approaches for caption generation. To process multimodal inputs, we first employ CLIP [40] to extract the image features that are dimensionally consistent with text embeddings. Subsequently, the image feature is incorporated as a distinct token at the beginning of the text embedding sequence, allowing for the processing of multimodal inputs through a unified language model framework. Upon finetuning, the models are trained to discern the semantic essence of the image feature, enhancing their capability to generate text outputs aligned with the given image.

Incorporating additional trainable modules to handle highlight inputs can escalate computation complexity and potentially hinder the utilization of the formidable capabilities of pretrained LLMs. We propose two simple yet effective methods: the Prompting-based Controller ($\mathcal{P}\text{-Ctrl}$) and Recalibration-based Controller ($\mathcal{R}\text{-Ctrl}$), enabling the model to yield controllable outputs based on highlighted context without modifying model architectures. This adaptation facilitates the easy adaptation of modern LLMs into the Ctrl-CIC task by finetuning them on the controllable data.

Prompting-Based Controller Incorporating the Prompting-based Controller ($\mathcal{P}\text{-Ctrl}$) allows for conditional caption generation by leveraging highlights as a prompt during decoding. First, we derive the training highlights using word-level relevance scores:

$$H = \{w_j \mid s_{w_j} > \theta, w_j \subseteq C\}, \quad (6)$$

where θ is the threshold value for highlight selection. Subsequently, we assemble a prompt string with elements $w_j \in H$, delineated by the special token $\langle \text{SEP} \rangle$. To steer the model towards conditional token generation based on the highlights, we affix this prompt string to the intended caption, as illustrated in Fig. 2b. We then finetune any modern LLM to predict these augmented texts based on the image and its context. The autoregressive nature of LLM enables them to condition the generation of captions based on the prefixed string, effectively learning to customize outputs in response to the given highlight prompt. During Ctrl-CIC inference, the prompt, constructed from test highlight inputs, is fed into the model as decoder inputs. This input prefix guides the model to produce captions specifically tailored to the provided highlights.

Recalibration-Based Controller Alternatively, a Ctrl-CIC model can be directly trained using token-level relevance scores with the Recalibration-based Controller ($\mathcal{R}\text{-Ctrl}$), as depicted in Fig. 2b. Firstly, we normalize the cosine similarity scores to weights within unit intervals:

$$\mathcal{W} = \{w_i = \frac{s_i}{2} + 0.5 \mid s_i \in \mathcal{S}\}.$$

During training, we perform element-wise multiplication for each of the encoder token embeddings with their token weight to $e'_i = e_i * w_i$. Through finetuning, the model is conditioned to utilize these weights that calibrate the feature distribution for guided text generation. For inference purposes, we further train a weight predictor based on an additional LM encoder, the encoded embeddings produced by which are then converted to weights \hat{e} within unit intervals with a linear and a sigmoid layer. To control caption generation, the model recalibrates the predicted weights by incrementing the token weights at i^{th} position by a value of α if that token is highlighted, *i.e.*, $\hat{e}'_i = \hat{e}_i \cdot (w_i + \alpha)$, thereby enhancing focus on the highlighted content.

Adaptability to CIC In addition to generating controlled captions pertinent to highlights, both proposed controllers are seamlessly adaptable for traditional CIC tasks. The $\mathcal{P}\text{-Ctrl}$ achieves this by generating captions based on a self-predicted prefix, without the need for constructing a highlights-based prompt. Meanwhile, the $\mathcal{R}\text{-Ctrl}$ accomplishes traditional CIC caption generation by adhering to the predicted weights \hat{e} , omitting any recalibration. The versatility of these models in performing traditional CIC tasks is further elaborated in the supplementary materials.

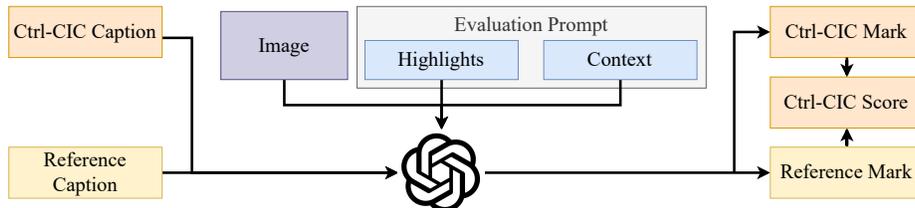


Fig. 3: GPT-4V empowered evaluator for Ctrl-CIC task. Given a pair of Ctrl-CIC and reference captions, this GPT-4V evaluator comprehensively reasons and marks them, and derives the final score as the ratio of raw marks between Ctrl-CIC and the reference caption. Note: the reference caption also serves as GT for the standard CIC task.

4 GPT-4V based Evaluation

As the highlights data used to train the model are constructed with a computational method rather than being curated by human expertise in real-world scenarios, we cannot entrust them as reliable references to evaluate the Ctrl-CIC performance of our model. Instead, we established an AI-enhanced evaluation framework, capitalizing on the unparalleled proficiency of the state-of-the-art language model, GPT-4V, to assess the controllability of Ctrl-CIC models.

Highlights Selection Our Ctrl-CIC models are designed to generate distinct captions for different sets of highlight inputs. To facilitate this, we employed GPT-4 to select a set of N candidate highlights, consisting of words or phrases h_1, h_2, \dots, h_N , from each data sample. To ensure the relevance and non-redundancy of these highlights within the context, we refined the set by removing highlights that overlap or are not found in the context. This refinement process ensures the integrity of our test sets, which are constructed to include varying numbers of highlight segments per sample.

Evaluation Inspired by [27], we adopt a comparative evaluation approach to measure the quality of the Ctrl-CIC captions. As illustrated in Fig. 3, we provide the GPT-4-Vision (GPT-4V) evaluator with the image, necessary context inputs, and corresponding highlights. GPT-4V then comparatively scores two captions: the reference caption (ground truth for CIC), acting as a comparative anchor, and the Ctrl-CIC caption under evaluation. These are assessed using various metrics, with scores ranging from 1 to 5. As the ground truth for the CIC task, the reference caption inherently possesses perfect alignment with the image and its context. This characteristic makes it a suitable comparative anchor for evaluating Ctrl-CIC captions. The relative quality of a Ctrl-CIC caption is then determined by dividing its mark by the reference caption mark, resulting in a score that accurately reflects the Ctrl-CIC captions’ quality relative to standard contextual captions. To mitigate asymmetrical scoring bias inherent in division-based metrics, we apply a logarithmic transformation to individual scores from

all test samples before computing their average. This approach symmetrically normalizes reciprocal values. Post-averaging, we exponentiate the mean score to appropriately rescale the final results, thereby preserving the relative quality assessment.

By employing this comparative scoring method, a consistent evaluative benchmark is applied to different Ctrl-CIC captions. They are assessed against a common anchor, effectively reducing subjective variability in GPT-4V’s scoring. To address the positional bias inherent in the GPT-4V-based evaluators, as identified in [55], we randomly alternate the input order of the two captions. This approach, contrasting with individual caption scoring, ensures uniformity in evaluation and mitigates the subjective variance inherent in assessments produced by GPT-4V.

To enhance the robustness of our evaluation methodology, we utilized multi-step evaluation prompts, guiding GPT-4V to systematically score the Ctrl-CIC captions in a chain-of-thought (CoT) format. Specifically, we instruct GPT-4V to generate CoT evaluation steps following [28], based on the predefined evaluation criteria for different metrics. Moreover, GPT-4V is tasked with producing analytical reasoning for each caption and comparative reasoning prior to scoring them, thereby enhancing the score quality generated through an auto-regressive process. Finally, we synthesize the task description, evaluation criteria, evaluation steps, and Ctrl-CIC inputs to formulate the complete evaluation query.

5 Experiments and Results

5.1 Datasets and Implementation

We conduct our experiments on Wiki-Web2M [5], comprising two million Wikipedia pages, for contextual image captioning studies. We implement our methods with the LongT5-base [14] architecture for its balance of performance and memory efficiency, and utilize CLIP-large [40] for image feature extraction and T5-large [41] for calculating token-level relevance scores. We evaluate our models on the official test split provided by [5], with highlights sets constructed using GPT-4. We refer to this test set as Wiki-Web2M_{full}. The specifics of hyper-parameters and training setups are detailed in the supplementary materials.

5.2 Baselines and Metrics

Ctrl-CIC Annotations For better reference purposes, we alleviate the difficulty of obtaining golden standard labels for the Ctrl-CIC tasks by introducing two sets of proxy annotations. The first set comprises human-generated annotations for the traditional CIC task, referred to as CIC-GT, where these labels serve as a paradigm of optimal context relevance and image-caption consistency. The second set consists of Ctrl-CIC labels generated by GPT-4 through a combination of context, GRIT image caption, and highlight sets. This GPT-4 reference was obtained on a subset of 5,000 randomly chosen test samples, referred to as Wiki-Web2M_{5k}, to ensure a comprehensive evaluation. Together, the evaluation results of these two sets of labels are indicated in gray as reference benchmarks.

Baselines We compare the performance of our Ctrl-CIC models against the following baselines. Firstly, we adapted a pretrained LongT5 [14] CIC model, for the Ctrl-CIC task. This adaptation, named LongT5-Ext, selects sentences from the context based on their inclusion of the highlight, creating a highlight-centered context. This modified context enables the CIC model to produce captions aligned with both context and highlights, fulfilling their role as a Ctrl-CIC baseline. Additionally, we train a LongT5-Tune baseline by finetuning the LongT5 model on the extracted highlight-centered context for the Ctrl-CIC task. We also integrate the LLaVA-1.5 model [27] in our comparisons, providing a benchmark against the latest vision-language models and offering insights into the capabilities of our models relative to current state-of-the-art advancements.

Metrics We evaluate the quality of Ctrl-CIC captions using a set of reference-free metrics. We use recall (R) to evaluate the inclusion rate of the highlights in the Ctrl-CIC captions. To evaluate caption diversity, we employ Div-N (D-N) [3], measuring the proportion of distinct N-grams across five different captions generated for the same context-image pair with respect to varying highlights. In addition to the above conventional metrics, we also incorporate model-based scores. Specifically, CLIPScore (CS) [17] is utilized to assess the semantic alignment between the caption and the image. Furthermore, we introduce CLIPScore-Sent (CS-S), a metric that calculates the cosine similarity between the CLIP text embedding of the caption and the averaged embedding of context sentences containing the highlights. Consequently, CLIPScore-Sent reflects the extent to which the captions are relevant to the specified highlights within the context. These metrics evaluate the capability of the captions to adapt to different highlights while maintaining consistency with the image.

In terms of the GPT-4V evaluator, following [28], prompts incorporating various metrics were used to guide GPT-4V in assessing the Ctrl-CIC captions against different criteria: *Context Relevance (CR)*, *Highlight Relevance (HR)*, *Image Consistency (IC)*, and *Overall Quality (OQ)*. The exemplary prompts for GPT-4V-based evaluation are demonstrated in the supplementary materials.

5.3 Ctrl-CIC Results

Quantitative Analysis We evaluate the performance of Ctrl-CIC methods under different highlight settings. Tab. 1 presents the quantitative results on test samples with a singular highlighted phrase in the context, and Tab. 2 demonstrates the quantitative results when multiple highlighted phrases are present.

Under both settings, recall and CLIPScore-Sent have indicated that both proposed models possess significantly enhanced relevance to highlights. Meanwhile, the Div-1 and Div-2 scores prove the diversity of the proposed controllers is superior to the baselines. The results indicate effective controls have been integrated into the caption generation process. While both Long-T5-based methods show some controllability compared to the CIC GT, their implicit highlight incorporation via extraction makes fine-grained control challenging. Additionally,

Table 1: Evaluation on Ctrl-CIC with a *single* highlighted segment per sample. Results are evaluated using conventional metrics including Recall (R), Div-1 (D-1), Div-2 (D-2), and model-based metrics including CLIPScore (CS), and CLIPScore-Sent (CS-S).

Model	Setting	Wiki-Web2M _{full}					Wiki-Web2M _{5k}				
		R ↑	D-1 ↑	D-2 ↑	CS ↑	CS-S ↑	R ↑	D-1 ↑	D-2 ↑	CS ↑	CS-S ↑
CIC GT [5]	-	8.28	3.24	2.76	68.66	51.30	12.05	3.20	2.76	68.09	53.00
GPT-4 [33]	-	-	-	-	-	-	86.10	7.09	8.38	65.09	56.33
LongT5-Ext	Pretrained	18.45	5.65	5.59	64.28	57.99	15.28	6.23	6.35	64.40	57.14
LongT5-Tune	Finetune	17.62	6.06	5.83	63.35	57.48	14.44	6.49	6.45	63.24	56.47
LLaVA-1.5 [27]	Zero-shot	36.57	7.09	8.38	65.56	50.34	32.56	7.25	8.64	65.53	47.98
<i>Our Approach.</i>											
\mathcal{P} -Ctrl	Finetune	57.07	10.23	10.54	60.91	59.57	57.34	10.53	10.83	60.14	59.59
\mathcal{R} -Ctrl	Finetune	64.88	9.21	10.30	63.60	59.27	63.5	9.47	10.65	62.70	58.94

Table 2: Evaluation on Ctrl-CIC with *multiple* highlighted segments per sample.

Model	Setting	Wiki-Web2M _{full}					Wiki-Web2M _{5k}				
		R ↑	D-1 ↑	D-2 ↑	CS ↑	CS-S ↑	R ↑	D-1 ↑	D-2 ↑	CS ↑	CS-S ↑
CIC GT [5]	-	9.6	3.23	2.76	67.81	55.89	16.30	3.25	2.78	68.66	55.16
GPT-4 [33]	-	-	-	-	-	-	84.02	7.98	9.82	67.89	53.69
LongT5-Ext	Pretrained	13.48	5.57	5.53	64.47	60.30	23.94	5.58	5.58	64.97	61.00
LongT5-Tune	Finetune	12.68	5.92	5.81	63.86	59.80	22.23	5.93	5.82	64.01	60.47
LLaVA-1.5 [27]	Zero-shot	26.06	6.53	7.82	66.38	51.88	34.56	6.48	7.74	66.88	52.14
<i>Our Approach.</i>											
\mathcal{P} -Ctrl	Finetune	44.37	9.45	10.34	61.59	62.94	56.62	9.41	10.27	61.32	68.01
\mathcal{R} -Ctrl	Finetune	52.34	8.97	10.22	62.67	63.00	58.85	8.96	10.19	62.53	66.59

truncating context to emphasize highlights often excludes vital background information, hindering the model’s capacity for comprehensive captioning. On the other hand, our proposed controllers achieve fine-grained highlight incorporation and effectively utilize the whole context, leading to significantly stronger controllability and caption quality.

Nonetheless, the Ctrl-CIC models revealed some challenges in maintaining consistency with the image content according to the CLIPScore. This observation aligns with expectations, given that the Ctrl-CIC captions mainly focused on the highlighted elements, potentially leading to a narrowed or biased interpretation of the visual data. This underscores a vital direction for future enhancement.

Evaluation with GPT-4V The GPT-4V-based evaluation results, as detailed in Tab. 3, corroborate our findings. Notably, the \mathcal{R} -Ctrl controller not only excels in ‘Highlight Relevance’ and ‘Overall Quality’ but also demonstrates comparable performance in ‘Context Relevance’ to the CIC Ground Truth. This suggests that the \mathcal{R} -Ctrl method effectively maintains relevance to the overall context, not just the highlighted segments.

Although the GPT-4V evaluation largely agrees with subjective assessments in aspects like highlight relevance and image consistency, some discrepancies are noted. For instance, the GPT-4V evaluator assigns LLaVA-1.5 the highest

Table 3: Evaluating Ctrl-CIC with GPT-4V. Scores of CIC ground truth (CIC GT) that serve as the comparative anchor are always normalized to unit values. The comparative scores reflect relative performance; scores exceeding 1 signify superior performance compared to the CIC GT caption.

	CR \uparrow	HR \uparrow	IC \uparrow	OQ \uparrow
CIC GT [5]	1	1	1	1
GPT-4 [33]	1.5	2.81	0.98	2.00
LongT5-Ext	0.95	1.03	0.88	1.17
LongT5-Tune	0.86	0.96	0.84	1.06
LLaVA-1.5 [27]	0.98	1.27	1.05	1.26
<i>Our Approach.</i>				
\mathcal{P} -Ctrl	0.94	1.64	0.75	1.21
\mathcal{R} -Ctrl	1.01	1.73	0.71	1.3

score for image consistency, surpassing even the reference labels, contrary to what CLIPScore indicates. This difference could be attributable to GPT-4V’s use of contextual clues for image content assessment, as opposed to CLIPScore’s reliance solely on visual information. Despite these variances, a correlation analysis confirms that the GPT-4V-based evaluator usually aligns more closely with human judgment than the subjective metrics. Further details of this analysis are available in the supplementary materials.

Discussion While both proposed controllers exhibit commendable levels of controllability, our results indicate that the \mathcal{R} -Ctrl tends to produce captions more directly related to the highlights compared to the \mathcal{P} -Ctrl. Conversely, the \mathcal{P} -Ctrl generates a greater diversity of captions for different highlights within the same context-image pair. This distinction likely arises because the \mathcal{R} -Ctrl pushes generation towards highlight-related tokens at each decoding step, which may limit its creative scope. In contrast, the \mathcal{P} -Ctrl implicitly conditions its controlled captions on the highlights implicitly, primarily through decoder attention, thus affording greater versatility in the caption generation process. Overall, our methods demonstrate superior performance compared to LLaVA-1.5 at generating controlled captions, with a mere 5% of its parameter count.

Qualitative Analysis Figure 4 shows some qualitative Ctrl-CIC samples generated by the \mathcal{P} -Ctrl model. It can be observed that our methods can tune the model to generate diverse captions for varied images relevant to different highlights. The model also exhibits a notable ability to extract relevant contextual hints to enrich the captions. For instance, in the ‘Mamamoo’ example, the model adeptly references the correct year associated with the highlighted event. While the generated captions are generally brief and feature simple linguistic structures, this reflects the style of our training data, which primarily consists of concise Wikipedia captions. Observations reveal similar results from the \mathcal{R} -Ctrl method, which are not included in the figures due to their resemblance.



Fig. 4: Qualitative demonstration on our Ctrl-CIC results produced by \mathcal{P} -Ctrl. Highlights and their respective Ctrl-CIC captions are aligned in colors, showing *how captions vary with different input images and highlights for the same context*. Section titles, if any, are appended after the page title. Paragraphs in the context that are without any highlights are omitted for readability.

Limitations The weakly supervised training approach may limit the generalization to real-life highlights and introduce biases. Incorporating synthesized highlights in training could help mitigate this. Furthermore, the \mathcal{R} -Ctrl method's uniform weight recalibration across all tokens can lead to repetitive generation of highlights. An adaptive recalibration strategy that modifies weight adjustment strength based on already generated tokens could address this issue.

6 Conclusion

In this work, we introduce a new challenging captioning task termed as Controllable Contextualized Image Captioning (Ctrl-CIC). Ctrl-CIC augments standard CIC by controlling over highlight elements, facilitating caption generation more attuned to user-specific intent. As initial solutions, we present \mathcal{P} -Ctrl and \mathcal{R} -Ctrl, two versatile approaches compatible with arbitrary LMs, enabling them to handle Ctrl-CIC without compromising performance in conventional CIC tasks. We further design a GPT-4V-based evaluation framework to assess the quality of the generated Ctrl-CIC captions from multiple dimensions. Empirical findings underscore the efficacy of our proposed methods, manifesting notable enhancements in controllability relative to existing captioning methods while maintaining superior efficiency.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 35, pp. 23716–23736 (2022)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition (CVPR)*. pp. 6077–6086 (2018)
3. Aneja, J., Agrawal, H., Batra, D., Schwing, A.G.: Sequential latent spaces for modeling the intention during diverse image captioning. In: *Proceedings of the IEEE international conference on computer vision (ICCV)*. pp. 4260–4269 (2019)
4. Biten, A., Gomez, L., Rusinol, M., Karatzas, D.: Good news, everyone! context driven entity-aware captioning for news images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 12458–12467 (2019)
5. Burns, A., Srinivasan, K., Ainslie, J., Brown, G., Plummer, B.A., Saenko, K., Ni, J., Guo, M.: A suite of generative tasks for multi-level multimodal webpage understanding. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2023)
6. Carlsson, F., Öhman, J., Liu, F., Verlinden, S., Nivre, J., Sahlgren, M.: Fine-grained controllable text generation using non-residual prompting. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. pp. 6837–6857 (2022)
7. Chen, H., Yi, X., Sun, M., Li, W., Yang, C., Guo, Z.: Sentiment-controllable chinese poetry generation. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*. pp. 4925–4931 (2019)
8. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022)
9. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020)
10. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500* (2023)
11. Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., Liu, R.: Plug and play language models: A simple approach to controlled text generation. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2020)
12. Ding, N., Deng, C., Tan, M., Du, Q., Ge, Z., Wu, Q.: Image captioning with controllable and adaptive length levels. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023)
13. Fei, J., Wang, T., Zhang, J., He, Z., Wang, C., Zheng, F.: Transferable decoding with visual entities for zero-shot image captioning. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 3136–3146 (2023)
14. Guo, M., Ainslie, J., Uthus, D., Ontanon, S., Ni, J., Sung, Y.H., Yang, Y.: LongT5: Efficient text-to-text transformer for long sequences. In: *North American Chapter of the Association for Computational Linguistics (NAACL)*. pp. 724–736 (2022)

15. Gupta, A., Jain, R.: Visual information retrieval. *Commun. ACM* **40**(5), 70–79 (1997)
16. He, X.: Parallel refinements for lexically constrained text generation with BART. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 8653–8666 (2021)
17. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2022)
18. Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., Wang, L.: Scaling up vision-language pre-training for image captioning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 17980–17989 (2022)
19. Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., Liu, Q., Aggarwal, K., Chi, Z., Bjorck, N., Chaudhary, V., Som, S., SONG, X., Wei, F.: Language is not all you need: Aligning perception with language models. In: Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 36, pp. 72096–72109 (2023)
20. Kalarani, A.R., Bhattacharyya, P., Chhaya, N., Shekhar, S.: “Let’s not Quote out of Context”: Unified vision-language pretraining for context assisted image captioning. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* (2023)
21. Khalifa, M., Elsahar, H., Dymetman, M.: A distributional approach to controlled text generation. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2021)
22. Koh, J.Y., Salakhutdinov, R., Fried, D.: Grounding language models to images for multimodal inputs and outputs. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2023)
23. Krause, J., Johnson, J., Krishna, R., Fei-Fei, L.: A hierarchical approach for generating descriptive image paragraphs. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. pp. 317–325 (2017)
24. Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., Zhang, J., Huang, S., Huang, F., Zhou, J., Si, L.: mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 7241–7259 (2022)
25. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: *Proceedings of the International Conference on Machine Learning (ICML)* (2023)
26. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: *Text Summarization Branches Out*. pp. 74–81 (2004)
27. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 36, pp. 34892–34916 (2023)
28. Liu, Y., Iyer, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-Eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634* (2023)
29. Liu, Y., Jia, Q., Zhu, K.: Length control in abstractive summarization by pretraining information selection. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. pp. 6885–6895 (2022)
30. Makino, T., Iwakura, T., Takamura, H., Okumura, M.: Global optimization under length constraint for neural text summarization. In: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. pp. 1039–1048 (2019)

31. Nguyen, K., Biten, A.F., Mafra, A., Gomez, L., Karatzas, D.: Show, interpret and tell: Entity-aware contextualised image captioning in wikipedia. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 1940–1948 (2022)
32. Nguyen, V.Q., Suganuma, M., Okatani, T.: GRIT: Faster and better image captioning transformer using dual visual features. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 167–184. Springer (2022)
33. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
34. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
35. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
36. Papineni, K., Roukos, S., Ward, T., Jing Zhu, W.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL). pp. 311–318 (2002)
37. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision (ICCV). pp. 2641–2649 (2015)
38. Puduppully, R., Dong, L., Lapata, M.: Data-to-text generation with content selection and planning. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). pp. 6908–6915 (2019)
39. Qu, T., Tuytelaars, T., Moens, M.F.: Visually-aware context modeling for news image captioning. arXiv preprint arXiv:2308.08325 (2023)
40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
41. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* **21**(140), 1–67 (2020)
42. Ramisa, A., Yan, F., Moreno-Noguer, F., Mikolajczyk, K.: Breakingnews: Article annotation by image and text processing. In: Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI). p. 1072–1085 (2018)
43. Ribeiro, L.F.R., Zhang, Y., Gurevych, I.: Structural adapters in pretrained language models for AMR-to-Text generation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 4269–4282 (2021)
44. Ruan, Y., Ling, Z.: Emotion-regularized conditional variational autoencoder for emotional response generation. *IEEE Transactions on Affective Computing* **14**(01), 842–848 (2023)
45. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics. pp. 2556–2565 (2018)
46. Song, H., Wang, Y., Zhang, K., Zhang, W.N., Liu, T.: BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 167–177 (2021)

47. Song, H., Wang, Y., Zhang, W.N., Liu, X., Liu, T.: Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 5821–5831 (2020)
48. Song, Z., Zheng, X., Liu, L., Xu, M., Huang, X.: Generating responses with a specific emotion in dialog. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 3685–3695 (2019)
49. Su, Y., Lan, T., Liu, Y., Liu, F., Yogatama, D., Wang, Y., Kong, L., Collier, N.: Language models can see: Plugging visual controls in text generation. arXiv preprint arXiv:2205.02655 (2022)
50. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
51. Tran, A., Mathews, A., Xie, L.: Transform and tell: Entity-aware news image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13035–13045 (2020)
52. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S.M.A., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 200–212 (2021)
53. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: GIT: A generative image-to-text transformer for vision and language. arXiv preprint arXiv:2205.14100 (2022)
54. Wang, N., Xie, J., Wu, J., Jia, M., Li, L.: Controllable image captioning via prompting. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 37, pp. 2617–2625 (2023)
55. Wang, P., Li, L., Chen, L., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., Sui, Z.: Large language models are not fair evaluators. arXiv preprint arXiv:2305.17926 (2023)
56. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 23318–23340 (2022)
57. Wang, Y., Xu, N., Tian, H., Lv, B., Duan, Y., Li, X., Liu, A.A.: Knowledge prompt makes composed pre-trained models zero-shot news captioner. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME). pp. 28779–2884 (2023)
58. Yang, P., Li, L., Luo, F., Liu, T., Sun, X.: Enhancing topic-to-essay generation with external commonsense knowledge. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL). pp. 2002–2012 (2019)
59. Yang, X., Karaman, S., Tetreault, J., Jaimes, A.: Journalistic guidelines aware news image captioning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 5162–5175 (2021)
60. Zeng, Z., Zhang, H., Lu, R., Wang, D., Chen, B., Wang, Z.: ConZIC: Controllable zero-shot image captioning by sampling-based polishing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23465–23476 (2023)
61. Zhang, J., Fang, S., Mao, Z., Zhang, Z., Zhang, Y.: Fine-tuning with multi-modal entity prompts for news image captioning. In: Proceedings of the ACM International Conference on Multimedia (MM). p. 4365–4373. MM '22 (2022)
62. Zhao, W., Wu, X.: Boosting entity-aware image captioning with multi-modal knowledge graph. IEEE Transactions on Multimedia pp. 1–12 (2023)

63. Zhao, Y., Liu, Y., Guo, Z., Wu, W., Gong, C., Wan, F., Ye, Q.: Controllable dense captioner with multimodal embedding bridging. arXiv preprint arXiv:2401.17910 (2024)
64. Zhong, P., Zhang, C., Wang, H., Liu, Y., Miao, C.: Towards persona-based empathetic conversational models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6556–6566 (2020)
65. Zhou, M., Luo, G., Rohrbach, A., Yu, Z.: Focus! relevant and sufficient context selection for news image captioning. In: Findings of the Association for Computational Linguistics: Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6078–6088 (2022)
66. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)