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

Supplementary Material

1 Datasets and Implementation Details

Dataset Details. We conduct our experiments on Wiki-Web2M [2], a large-scale webpage dataset containing two million Wikipedia article pages with text and images organized by sections, allowing context to be defined easily. Specifically, we follow the official split on the contextual image captioning task, with 2,222,814/124,703/124,188 samples for train/val/test split respectively. Each sample consists of an image paired with its relevant context derived from the webpage data. The structured data allows for a more flexible definition of context at different levels. In the presented experiment, the context of images is constructed by combining the page title, section title, section text, and captions of the remaining images in the section of the webpage.

Model Details. We adopt LongT5 (tglobal-base) [4] as the underlying architecture for both proposed methods as well as the weight predictor for $\mathcal{R}$-`Ctrl`, due to its superior efficiency compared to Prefix Global used in [2] with only a marginal trade-off in performance. We use CLIP-large [12] to extract the frozen image feature, and we adopt T5-large [13] to derive the token-level relevance scores for training.

Hyperparameter Details. Input texts are truncated to a length of 512 tokens, and the output length is set to 128 for all methods except $\mathcal{P}$-`Ctrl`, where an output length of 192 is used in account for the extra prompting tokens. For $\mathcal{P}$-`Ctrl`, the threshold value to derive the training highlights prompt is set to $\theta = 0.3$. Furthermore, we impose an upper limit of 40 on the number of prompting words during training to prevent scenarios where an excessively lengthy prompt could constrain the space available for the caption. For $\mathcal{R}$-`Ctrl`, we set the weight recalibration value to $\alpha = 0.1$ for controlled caption generation. During training, all methods are optimized with an AdamW [8] optimizer with a linearly decaying learning rate of $5e-5$ and betas $(0.9, 0.999)$. The models are trained on a single NVIDIA RTX 3090 GPU, with a batch size of 12 for 3 million steps. In addition, the weight predictor for $\mathcal{R}$-`Ctrl` is trained using cross entropy loss with the same optimizer and a batch size of 20 for 1 million steps.

GPT-4V Evaluation Details. In terms of highlights selection, we prompt GPT-4 to select $N = 10$ highlighted component candidates from contexts, and

Table 1: Cost analysis for Ctrl-CIC models. The per-sample inference time are estimated based on the average query expenses. The LongT5-Tune method, having identical architecture and inputs as LongT5-Ext, is omitted from the table. The parameter count for GPT-4, unofficially reported as 8 × 220 billion, remains unconfirmed. Notably, besides the vast parameter size and time cost, GPT-4 also poses a monetary cost of approximately $0.0059 per Ctrl-CIC query.

Model	Parameters (B)	Time (s)
GPT-4	8×220 †	0.563
LongT5-Ext	0.248	0.169
LLaVA-1.5	7.063	0.634
$\mathcal{P}$-`Ctrl`	0.248	0.166
$\mathcal{R}$-`Ctrl`	0.357	0.223

Table 2: Correlation analysis between human judgment and different metrics. The 'G-' prefixed rows represent the results from GPT-4V evaluation, whereas the 'H-' prefixed column headings denote human assessments. 'CS' and 'CS-S' stand for CLIPScore and CLIPScore-Sent respectively. The metric owning best human alignment, is highlighted in **bold**, and the second-best one is <u>underlined</u>.

	H-CR			H-HR			H-IC			H-OQ		
	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
CS	0.479	0.476	0.344	-0.198	-0.228	-0.171	0.546	0.571	0.438	0.231	0.198	0.156
CS-S	0.491	0.511	0.375	0.241	<u>0.323</u>	0.233	0.336	0.183	0.063	0.501	0.614	0.469
Recall	-0.032	0.129	0.112	**0.401**	0.309	<u>0.267</u>	-0.200	-0.207	-0.180	0.131	0.258	0.225
G-CR	**0.665**	**0.791**	**0.639**	0.217	0.232	0.150	0.547	0.747	<u>0.589</u>	0.570	0.573	0.488
G-HR	0.476	0.691	0.560	<u>0.381</u>	**0.495**	**0.317**	0.277	0.351	0.336	<u>0.572</u>	**0.734**	**0.560**
G-IC	<u>0.629</u>	0.694	0.548	-0.001	0.012	0.032	**0.690**	**0.778**	0.581	0.410	0.382	0.323
G-OQ	0.613	<u>0.760</u>	<u>0.606</u>	0.224	0.264	0.163	<u>0.594</u>	<u>0.769</u>	**0.623**	**0.594**	<u>0.626</u>	<u>0.525</u>

construct the evaluation sets with different numbers of highlights based on the filtered highlights. Contexts with insufficient numbers of selected highlights are also filtered during constructing *single* and *multiple* highlights test sets. For GPT-4V empowered evaluation, we randomly select 50 samples from the test set, on which the GPT-4V is prompted to evaluate the Ctrl-CIC captions from different methods.

2 Model Cost Analysis

Table 1 presents the cost analysis in terms of parameter size and inference time for Ctrl-CIC models. For fair comparisons, input contexts across all methods are uniformly truncated to the same length. With a common LongT5 backbone architecture, both controller methods enhance controllability relative to LongT5-Ext, while maintaining similar levels of efficiency. $\mathcal{R}$-`Ctrl` exhibits marginally lower efficiency than $\mathcal{P}$-`Ctrl` due to its additional weight predictor. Nonetheless, both controllers substantially outperform GPT-4 and LLaVA-1.5 in terms

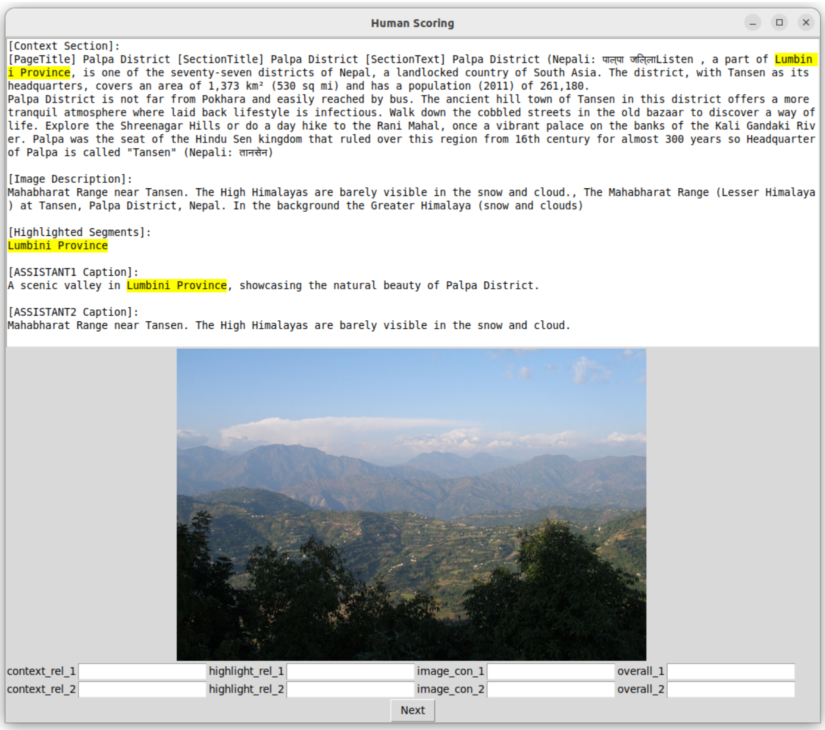

Fig. 1: UI of the scoring application for human evaluation. The top text box displays the context, image description, and the highlighted segments, along with the two captions to be scored based on Context Relevance, Highlight Relevance, Image Consistency, and Overall Quality. Human markers will review the context and score both the reference and Ctrl-CIC captions regarding each metric respectively.

of memory and computational efficiency. Notably, while GPT-4 generates high-quality Ctrl-CIC captions fairly efficiently, its high monetary costs limit its practicality in various real-world scenarios. In conclusion, our methods offer more efficient and cost-effective solutions for the Ctrl-CIC task.

3 Correlation Analysis of GPT-4V Evaluation

We conduct human evaluations on Ctrl-CIC captions to validate the alignment between the GPT-4V empowered evaluator and human judgment. using the same four metrics: *Context Relevance (CR)*, *Highlight Relevance (HR)*, *Image Consistency (IC)*, and *Overall Quality (OQ)*. As depicted in Fig. 1, we present the contextual information and anonymous captions to human markers and ask them to score the captions based on each metric. The human scoring of the Ctrl-CIC caption is then calculated based on the relative ratio of its raw mark to the reference caption's mark.

We conduct correlation analysis to examine the agreement between human judgment and metrics used in the paper. As shown in Tab. 2, Pearson (r), Spearman (ρ), and Kendall's tau (τ) coefficients are utilized to determine the strength of association between human judgment and metric scores. Div-N are not included in the analysis as they aim to evaluate the diversity between a set of captions on the same context-image pair, and are irrelevant to qualitative aspects as measured by other metrics.

It is notable that the metrics generated by the GPT-4V empowered evaluator frequently exhibit the highest correlations when compared to their human-judged counterparts, as evidenced by the corresponding pairs in the analysis (*e.g.*, 'G-CR' compared to 'H-CR'). For example, in assessing image-text consistency, 'G-IC' shows a higher correlation with 'H-IC' compared to CLIPScore. This higher correlation with human judgment may stem from GPT-4V's ability to utilize contextual clues in evaluating image content, in contrast to CLIPScore, which relies exclusively on visual information.

We also find that 'G-HR' shows correlation levels with 'H-OQ' that are akin to those observed for 'G-OQ'. This is attributed to the different evaluative focuses: human annotators value highlight relevance more in their assessments, while the GPT-4V empowered evaluator prioritizes context relevance and image consistency. This observation is further corroborated by examining the correlation between 'G-OQ' and human judgment across various aspects, where it is noted that 'H-HR' has the least correlation with 'G-OQ'.

Furthermore, the correlation analysis reveals an intriguing challenge in Ctrl-CIC captions: achieving a close relationship with highlights while also maintaining high consistency with the image content. Metrics that evaluate the relevance of captions to highlights often show zero or negative correlations with those assessing caption-image consistency. This finding supports our observations on the Ctrl-CIC evaluation results, where the proposed controllers, being more focused on highlights, tend to exhibit lower image consistency.

4 Adaptability of Ctrl-CIC Models to Standard CIC Task

In this subsection, we explore the adaptability and robustness of our Ctrl-CIC models when applied to the traditional CIC task. As previously noted, $\mathcal{P}$-`Ctrl` is capable of autonomously generating CIC captions based on the prefix prompts generated during the auto-regressive decoding, while our $\mathcal{R}$-`Ctrl` can be effectively adapted to the CIC task by maintaining the predicted weights without any recalibration. This section presents the experimental results of these adaptations, showcasing the flexibility and efficacy of our models in transitioning between Ctrl-CIC and CIC tasks.

Metrics. Following Wiki-Web2M [2], we report BLEU-4 (B-4) [11] , RougeL (R-L) [6], CIDEr (C) [15]. we also compute Meteor (M) [1] for CIC evaluation.

Table 3: Evaluation on standard CIC task on the test split of Wiki-Web2M benchmark (Wiki-Web2M$_{\text{test}}$). GPT-4 CIC results are obtained and reported on a 1,000 sample subset (Wiki-Web2M$_{1k}$). * denotes an extra input token length of 1,024 used. GPT-4 ($\cdot$) denotes the zero-shot inference of GPT-4 with different inputs.

		Wiki-Web2M$_{\text{test}}$				Wiki-Web2M$_{1k}$			
Model	Setting	B-4 ↑	R-L ↑	C ↑	M ↑	B-4 ↑	R-L ↑	C ↑	M ↑
Prefix Global* [2]	Finetune	10.92	36.21	148.53	-	-	-	-	-
LongT5 [4]	Finetune	10.31	36.00	142.77	30.58	9.42	35.01	128.03	29.88
InstructBlip [3]	Zero-shot	1.40	14.03	12.91	16.36	1.39	13.70	14.97	15.93
LLaVA-1.5 [7]	Zero-shot	2.55	16.48	25.07	16.81	2.71	16.06	23.63	16.40
GPT-4 [10] (w/ GRIT)	Zero-shot	-	-	-	-	4.81	21.55	40.05	23.13
GPT-4 [10] (w/ Attri)	Zero-shot	-	-	-	-	9.29	32.54	65.52	40.48
Our Approach.									
$\mathcal{P}$-`Ctrl`	Finetune	10.88	34.48	130.80	30.78	10.24	33.37	119.94	29.98
$\mathcal{R}$-`Ctrl`	Finetune	10.31	34.58	130.12	30.83	10.08	33.15	117.56	29.74

Baselines. Our primary baseline of comparison is the Prefix Global method [2], which stands as the sole existing CIC benchmark on the Wiki-Web2M dataset [2]. Owing to computational constraints, we reference results directly from [2]. For a more computation-friendly benchmark, we adapted the [2] baseline employing LongT5 [4]. Similar to Ctrl-CIC, we benchmark the zero-shot ability of InstructBlip [5], LLaVA-1.5 [7], and GPT-4 [10] on the CIC task. For GPT-4, we report results on a subset of 1,000 samples, referred to as Wiki-Web2M$_{1k}$, with two types of text-based image descriptions: image captions produced by GRIT [9], and the image attribution text from the dataset.

Results. The performance of models on the CIC task is shown in Tab. 3. Compared to the LongT5 baseline, both $\mathcal{P}$-`Ctrl` and $\mathcal{R}$-`Ctrl` exhibit higher or comparable scores in B-4 and Meteor metrics, indicating our proposed approaches are effective in generating high-quality contextual captions for the CIC task. Although our models do not reach the performance levels of the Prefix Global method, this is understandable given the computational constraints that limit both our models and the LongT5 baseline to shorter input lengths. The Prefix Global architecture has been shown to be more effective but also more computationally intensive than the LongT5 architecture [2]. In contrast, zero-shot models demonstrate less competitiveness in the CIC task. InstructBlip, for instance, struggles to generate coherent captions due to its limitations in processing extensive context. Other CIC captions generated by zero-shot models show notable improvement over InstructBlip, producing decent captions but still not reaching the accuracy level of fine-tuned models. The evaluation results demonstrate the seamless adaptability of the proposed Ctrl-CIC controllers to the standard CIC task, with minimal sacrifice on the performance.

Table 4: Llama-2 quantitative results in a pure textual-context-based Ctrl-CIC formulation.

		Wiki-Web2M$_{full}$					Wiki-Web2M$_{5k}$				
Model	Split	R ↑	D-1 ↑	D-2 ↑	CS ↑	CS-S ↑	R ↑	D-1 ↑	D-2 ↑	CS ↑	CS-S ↑
CIC GT [2]	Single	8.28	3.24	2.76	68.66	51.30	12.05	3.20	2.76	68.09	53.00
Llama-2 [14]	Single	65.26	10.11	12.35	53.12	65.97	61.36	10.15	12.42	52.63	65.93
CIC GT [2]	Multiple	9.6	3.23	2.76	67.81	55.89	16.30	3.25	2.78	68.66	55.16
Llama-2 [14]	Multiple	46.25	9.77	11.96	51.96	68.06	46.30	9.88	11.92	51.68	68.30

Table 5: Llama-2 evaluation results with GPT-4V.

	CR ↑	HR ↑	IC ↑	OQ ↑
CIC GT [2]	1	1	1	1
Llama-2 [14]	1.28	2.08	0.61	1.10

5 Analysis on the Importance of Visual Inputs

Some images in Ctrl-CIC simply depict the primary object in context, implying that textual context alone might suffice for generating adequate Ctrl-CIC captions and downplay the importance of visual inputs. However, it is more commonly observed that images contribute essential supplementary information that enhances understanding beyond mere illustration of the context, underscoring the vital role of visual inputs in the Ctrl-CIC tasks. Quantitatively, to demonstrate the significance of visual inputs, we provide a zero-shot evaluation using Llama-2 as a pure language model (LLM) baseline without any visual or image description inputs, as detailed in Tabs. 4 and 5. Observations show that employing an LM-based model that relies solely on textual context for Ctrl-CIC, while enhancing relevance with highlights and context, reduces image consistency and overall caption quality.

6 Qualitative Analysis on Ctrl-CIC

In this section, we demonstrate qualitative samples on the Ctrl-CIC task to showcase the capabilities of proposed models at Ctrl-CIC. The samples are shown in Figs. 2 to 5, where highlights and their respective Ctrl-CIC captions are aligned in colors. It can be observed that each model endeavors to infer the image caption with a particular focus on the elements highlighted within the image.

7 GPT-4 Prompts for Highlight Selection

Table 6 demonstrates an example prompt and response from GPT-4 for selecting the highlight from context to construct evaluation samples.

8 GPT-4 Prompts for Ctrl-CIC

Table 7 demonstrates an example prompt and response from GPT-4 for performing the Ctrl-CIC task.

9 GPT-4V Prompts for Ctrl-CIC Evaluation

Tables 8 to 10 demonstrate an example prompt and response from GPT-4V for comparatively evaluating two Ctrl-CIC captions. Note that images will also be given to the GPT-4V-empowered evaluator.

10 GPT-4 Prompts for CIC

Table 11 demonstrates an example prompt and response from GPT-4 for performing the CIC task.

Fort Knox Gold Mine -- Mining & milling

In 2003, the mine was scheduled to excavate up to 130,000 tonnes (140,000 short tons) per day with a **2.3:1 stripping ratio**, operating seven days per week twelve months per year. Due to its very **low sulfide content**, waste rock from Fort Knox does not produce acid mine drainage. Ore from the mine is processed at the Fort Knox mill, a conventional crushing, grinding, CIL leaching, carbon stripping and electrolytic recovery process, with a capacity of 36,000 tonnes (40,000 short tons) per day. The mill consists of a gyratory crusher, a 34' x 15' semi-autogenous grinding mill, two 20' x 30' ball mills, and a gravity separation circuit with a **Knelson concentrator**. Gold not recovered by the Knelson concentrator is sent through the CIL circuit.

In 2009, a valley-fill heap leaching operation came on line. Lower grade ore is trucked to the facility from the mine.

CIC Caption: Semi-autogenous grinding mill (foreground) at Fort Knox
GPT-4:
Large machine operating at a 2.3:1 stripping ratio inside the Fort Knox mill.
Large machine processing ore with low sulfide content at the Fort Knox mill
Large machine inside the Fort Knox mill, featuring the Knelson concentrator for gold separation.

P-Ctrl:	**R-Ctrl:**
The Knelson concentrator, which has a 2.3:1 stripping ratio	**Fort Knox mill - 2.3:1 stripping ratio**
The mine is very low in sulfide content	**Fort Knox mill with low sulfide content**
Knelson concentrator in operation	**Knelson concentrator at the Fort Knox mill**

Fig. 2: Ctrl-CIC Captions Sample A

Headlines (Friendship Never Ends) -- Live performances

In November 2007, the group performed together for the first time in nearly a decade at the **2007 Victoria's Secret Fashion Show**, held in Los Angeles, California. The group dressed in military-themed outfits performed their old hit single "Stop" miming to a backing track, in front of giant glittering lights that spelled out "Spice" in the background. Then they performed "Headlines", wearing gowns from Roberto Cavalli, but the performance was excluded from the TV broadcast for unknown reasons. A taped performance of the group lyp-synching the song aired on 17 November 2007 for the **Children in Need 2007 marathon**. For this performance the singers sang the song wearing 1930s long gowns, with Halliwell in a fuchsia satin number and Chisholm dressed in black.

"Headlines (Friendship Never Ends)" was also performed on their reunion tour, titled **The Return of the Spice Girls Tour** (2007–2008). Kitty Empire from The Observer reviewed the performance negatively, saying the song was "so unworthy of headlines that not even their fans have bought it". David Sinclair, whilst writing for The Sunday Times, also commented that "the pace flagged briefly with Headlines (Friendship Never Ends), the rather drab new song, which was accompanied by excerpts from its ill-judged promotional video". In 2019, Bunton confirmed to The Guardian that they would not perform the song on the Spice World – 2019 Tour.

CIC Caption:
The Spice Girls performing "Headlines (Friendship Never Ends)" during The Return of the Spice Girls Tour
GPT-4:
Two women performing on stage at the 2007 Victoria's Secret Fashion Show
Two women performing on stage during the Children in Need 2007 marathon
Two women performing on stage during The Return of the Spice Girls Tour

P-Ctrl:
The Spice Girls performing "Headlines" at the 2007 Victoria's Secret Fashion Show
The group performing "Headlines" during the Children in Need 2007 marathon
The Spice Girls performing "Headlines" on The Return of the Spice Girls Tour in 2008

R-Ctrl:
The Spice Girls performing "Headlines" at the 2007 Victoria's Secret Fashion Show
The Spice Girls performing "Headlines" during the Children in Need 2007 marathon.
The Spice Girls performing "Headlines" on The Return of the Spice Girls Tour, 2007

Fig. 3: Ctrl-CIC Captions Sample B

Yelena Mizulina -- Legislative work

Yelena Mizulina is among the lead authors of a set of **controversial legislative projects** including:

- 2010 bill on On Protecting Children from Information Harmful to Their Health and Development
- **2012 Internet Restriction Bill**.
- Future amendment to the bill On Protecting Children from Information Harmful to Their Health and Development that would block all websites containing swearing and the usage of mat. Opponents to the Internet Restriction Bill stressed that under the pretext of protecting children, this bill provides overwhelming tools for a broad scale Internet censorship and a limitation of freedom of speech in Russia. This protest has leading to public outcry such as the Russian-language Wikipedia strike. On June 10, 2012, she incriminated foreign intervention in these protests and therefore, she announced that she would issue an inquiry to the ministry of justice of the US. She also claimed that the Internet protests were organized by a "pedophile lobby".
- 2013 Anti-Magnitsky bill denying Americans the right to adopt Russian children. Following this proposal, US authorities have responded by proposing to include Y. Mizulina, along with Vitaly Milonov within the Magnitsky list of banned Russian personalities.
- 2014 bill preventing women from entering higher education before giving birth.
- 2017 bill to decriminalise domestic violence "first assaults which cause less serious injuries" to an administrative offence.

On November 14, 2012, Yelena Muzilina declared that the 'prophylactic goals' of Internet Restriction Bill of ' acquiring a safe information space on the web without making use of punitive measures' had been accomplished. She also disclosed a possible future legislative project that would prevent access to sites previously included in the **Russian Internet blacklist**. Among such portals are expected to fall rublacklist.net belonging to Pirate Party of Russia.

CIC Caption:
Yelena Mizulina talking to media on 11 July 2012 concerning the passing of the Internet Restriction Bill.
GPT-4:
Yelena Mizulina and her team discussing controversial legislative projects outside a building
Protesters gather outside a building to oppose the 2012 Internet Restriction Bill in Russia.
Protesters gather outside a building to oppose the Russian Internet blacklist legislation.

P-Ctrl:
Mizulina at a legislative meeting
Mizulina during the Internet Restriction Bill launch in 2012
Mizulina at the Internet blacklist protest in Moscow, 2012

R-Ctrl:
Yelena Mizulina during a legislative project on controversial issues, 2012
Yelena Mizulina at the 2012 Internet Restriction Bill launch
Yelena Mizulina at the Russian Internet blacklist press conference, 2012

Fig. 4: Ctrl-CIC Captions Sample C

Rib vault

A rib vault or ribbed vault is an **architectural feature** for covering a wide space, such as a **church nave**, composed of a framework of crossed or diagonal arched ribs. Variations were used in Roman architecture, Byzantine architecture, Islamic architecture, Romanesque architecture, and especially Gothic architecture. Thin stone panels fill the space between the ribs. This greatly reduced the weight and thus the outward thrust of the vault. The ribs transmit the load downward and outward to specific points, usually rows of columns or piers. This feature allowed architects of Gothic cathedrals to make higher and thinner walls and much larger windows.

It is a type of arcuated, or arched, vault in which the severies, or panels in the bays of the vault's underside are separated from one another by ribs which conceal the groins, or the intersections of the panels. Rib vaults are, like groin vaults, formed from two or three intersecting barrel vaults; the ribs conceal the junction of the vaults.

The earliest surviving example in Islamic architecture is at the **Mosque–Cathedral of Córdoba** in al-Andalus, which predates the earliest Romanesque examples by a century. An alternative to barrel vaults in the naves of churches, rib vaults in 12th century early Gothic architecture began to be used in vaults made with pointed arches, already known in the Romanesque style. In these vaults, as in groin vaults, the weight was directed it to the corners, where piers, columns, or walls could support it. Walls in Gothic buildings were often abutted by flying buttresses. These elements made it possible to construct buildings with much higher and thinner walls than before, with immense bays, and larger stained glass windows filling the structure with light.

CIC Caption:
Gothic rib vaulting, Reims Cathedral (begun 1221) nave
GPT-4:
Magnificent rib vault architecture in a large stone building
Arched ceiling of a church nave showcasing the rib vault architecture
Arched ceiling of the Mosque-Cathedral of Córdoba showcasing rib vault architecture

P-Ctrl:
The rib vault of the Church of Saint-Pierre and Saint-Paul in Paris, France, is a distinctive feature of the Gothic style.
Rib vault of the Church of Saint-Pierre and Saint-Paul, Paris
Rib vault of the Church of Saint-Pierre and Saint-Paul, in Paris

R-Ctrl:
Rib vault at the Cathedral of Saint-Pierre, France, a prominent feature of Gothic architecture
Rib vaults in the nave of the Church of Saint John the Baptist, Paris
Rib vault of the Mosque–Cathedral of Córdoba.

Fig. 5: Ctrl-CIC Captions Sample D

Table 6: GPT-4 prompt for selecting highlights for Ctrl-CIC evaluation

Prompt for Evaluation Highlights Selection

Task:

From the provided context, extract and highlight ten different informative keyphrases (or keywords) that contain valuable and relevant information. The extracted content can be a single word or phrases made up with several words. They should encapsulate key details, descriptive elements, or contextual insights that would be valuable for image captioning tasks. The aim is to ensure that some meaningful caption texts can be generated relevant to these context. Output these extracted keyphrases in one line, splitted by “|" (i.e., in a restricted format as “keyphrase1 | keyphrase2 | ..."; this requirement is very important and must be followed with no excuse) DIRECTLY (without any explanation or details).

Context Section:

Vermillion Township, Ashland County, Ohio. Vermillion Township, Ashland County, Ohio. Vermillion Township is one of the fifteen townships of Ashland County, Ohio, United States. The 2010 census found 2,618 people in the township, 2,170 of whom lived in the unincorporated portions of the township.

Keyphrase Extraction:

Response:

Vermillion Township | Ashland County | Ohio | fifteen townships | 2010 census | 2,618 people | unincorporated portions | United States | 2,170 residents | township population

Table 7: GPT-4 prompt for the Ctrl-CIC task

Prompt for the Ctrl-CIC Task
Controllable Contextual Image Captioning Task
Objective: Your goal is to generate a short image caption that aligns with the provided context, particularly focusing on the highlighted segments.

Instructions:

Context Section: Read the context provided to understand the overall theme or narrative. The context section will be given as the combination of a page title, a section title, and a section body. Some extra images other than the target image in the context section, if any, are included in the context section and described by [ImageCaptions] followed by text descriptions of the image to enrich the context.

Image Description: Analyze the brief description or attributes of the image.

Context Highlight: Pay special attention to the highlighted segments in the context. These are crucial cues for your caption. Highlights from different positions will be separated by line break.

Caption Generation: Construct a one-sentence caption for the image that is both descriptive of the image and relevant to the context, especially the highlighted context cues. Output the caption DIRECTLY (one sentence without any explanation or details).

Task Data:

Context Section:
[PageTitle] USS Vermont (BB-20) [SectionTitle] USS Vermont (BB-20) [SectionText] USS Vermont (BB-20), a Connecticut-class battleship, was the second ship of the United States Navy named after the 14th state. She was the third member of the class, which included five other ships. The Connecticut-class ships were armed with a main battery of four 12-inch (305 mm) guns and had a top speed of 19 knots (35 km/h; 22 mph). Vermont was laid down in May 1904 at the Fore River shipyard and launched in August 1905. The ship entered service with the Atlantic Fleet in March 1907.
Shortly after she entered service, Vermont joined the Great White Fleet for its circumnavigation of the globe in 1908–1909. She took part in the international Hudson–Fulton Celebration in New York in 1909 and made trips to Europe in 1910 and 1913. Thereafter, the ship became involved in interventions in several Central American countries, including the United States occupation of Veracruz during the Mexican Revolution, where two of her crew earned the Medal of Honor. During the United States' participation in World War I from April 1917 to November 1918, Vermont served as a training ship for engine room personnel. From November 1918 to June 1919, she made a series of trips to return American soldiers from Europe before being decommissioned in June 1920. She was sold for scrap in November 1923 according to the terms of the Washington Naval Treaty.

Image Description:
a ship with smoke coming out of it

Context Highlight:
Connecticut-class battleship

Controllable contextual caption (do not add quotations surrounding the response):

Response:
Connecticut-class battleship USS Vermont (BB-20) sailing with smoke billowing from its stacks.

Table 8: GPT-4V prompt for the Ctrl-CIC evaluation

Prompt for Ctrl-CIC Evaluation

We would like to request your feedback on the performance of two AI assistants in response to the controllable contextual image captioning task, where captions for an image will be generated based on the overall context and specific contextual highlights provided.

Evaluation Steps:
1. You will be given the image, [Context Section] and [Highlighted Segments], followed by controllable contextual captions provided by the two assistant [ASSISTANT1 Caption] and [ASSISTANT2 Caption]: The context section will be given as the combination of a page title, a section title, and a section body. The highlighted segments are parts of the section body which will be given as words, phrases, or sentences, separated by line breaks.
2. You will thoroughly read the [Context Section] and [Highlighted Segments] provided, and carefully examine the [Image].
3. You will read the caption generated by the AI assistant.
4. You will evaluate the controllable contextual image captioning quality of the two AI assistants, in terms of 4 aspects (which are "Relevance with Context", "Relevance with Highlight", "Consistency with Image", and "Overall Quality") - see below for individual criteria of these aspects. Each criterion should be considered in isolation to provide a clear and focused evaluation.
5. You will complete the following five sections IN ORDER (namely, [ASSISTANT1-Reasoning], [ASSISTANT2-Reasoning], [Comparison-Reasoning], [ASSISTANT1-Score], and [ASSISTANT2-Score])

Evaluation Criteria:
- Relevance with Context: This metric rates how relevant the caption is to the given context. It assesses whether the caption pertains to and is appropriate for the contextual information provided, without necessarily reflecting the entire context. Captions should be scored based on their pertinence and the degree to which they relate to the context. Annotators should deduct points for captions that do not relate to or ignore the context. Higher scores should be awarded to captions that show a clear and significant relevance to the context.
- Relevance with Highlight: This metric evaluates how well the caption aligns with the highlighted segments provided. The caption should accurately reflect the information contained in the highlighted segments, ensuring that it is relevant and integrated into the overall caption. Annotators are advised to penalize captions that fail to address the highlighted segments or do so in a manner that does not give them adequate prominence or relevance.
- Consistency with Image: This metric evaluates the accuracy with which the caption represents elements or themes that are verifiably present in the image, based on the provided image descriptions. The caption should not introduce content that is clearly absent from the image. It needs to maintain a clear and direct connection to the key elements depicted in the image. Annotators should deduct points for captions that include inconsistencies or introduce elements not discernible in the image. Higher scores should be reserved for captions that are faithful to the image's visible content.
- Overall Quality: This metric assesses the caption's overall effectiveness in the CCIC task, emphasizing its coherence with the overall context, alignment with the image, and relevance to the highlighted segment. A high-quality caption should seamlessly integrate these elements, providing an accurate, informative, and engaging description of the image that resonates with the given context and highlights.

To Be Continued Next Page

Table 9: GPT-4V prompt for Ctrl-CIC evaluation

HINT:
1. [ASSISTANT1-Reasoning] and [ASSISTANT2-Reasoning] will be used to record your reasoning and comments on the controllable contextual image caption generation quality of the two AI assistants, respectively;
2. [Comparison-Reasoning] will be used to record your feedback (with supporting evidence) for comparisions between the two AI assistants, which will be used to support the below two marking sections;
3. [ASSISTANT1-Score] and [ASSISTANT2-Score] will be used to record your controllable contextual image caption scores of the two AI assistants, respectively. Each assistant receives an integer score on a scale of 1 to 5 for each criteria, where a higher score indicates better performance according to the evaluation criteria.
4. Below is an example for requested output format of measuring one of the assistants:

[ASSISTANT1-Reasoning] (*example):
- Relevance with Context: some feedback with supporting evidance...
- Relevance with Highlight: some feedback with supporting evidance...
- Consistency with Image: some feedback with supporting evidance...
- Overall Quality: some feedback with supporting evidance...

[ASSISTANT1-Score] (*example):
- Relevance with Context: 3
- Relevance with Highlight: 1
- Consistency with Image: 4
- Overall: 2

-—————-Evaluation Starts——————

[Context Section]:
[PageTitle] Euphrates Handmade Syrian Horses and Riders [SectionTitle] Euphrates Handmade Syrian Horses and Riders [SectionText] The Euphrates Handmade Syrian Horses and Riders (EU_HSHRs) are zoomorphic clay figurines representing horses and horses with riders. They date from the late Iron Age period (mid 8th–7th centuries BCE)and were produced in the Middle Euphrates region, alongside anthropomorphic figures known as Euphrates Syrian Pillar Figurines (EU_SPFs).

[Highlighted Segments]:
zoomorphic clay figurines

[ASSISTANT1 Caption]:
EU_HSHR from Karkemish. The British Museum, London (museum no. 105006).

[ASSISTANT2 Caption]:
Euphrates Handmade Syrian Horses and Riders, zoomorphic clay figurines

Please make sure you read and understand these instructions carefully, and complete the following five sections IN ORDER: (1) firstly reasons via "[ASSISTANT1-Reasoning]:" and "[ASSISTANT2-Reasoning]:"; (2) secondly compares two assistants via "[Comparison-Reasoning]:"; and (3) finally marks them via "[ASSISTANT1-Score]:" and "[ASSISTANT2-Score]:".

To Be Continued Next Page

Table 10: GPT-4V response in the Ctrl-CIC evaluation

Response:
[ASSISTANT1-Reasoning]:
- Relevance with Context: The caption provided by Assistant 1 includes the acronym EU_HSHR, which is relevant to the context of Euphrates Handmade Syrian Horses and Riders. However, it does not mention the specific characteristic of being "zoomorphic clay figurines," which is a key aspect of the context.
- Relevance with Highlight: The caption fails to mention the highlighted segment "zoomorphic clay figurines," which is a significant oversight as this is a defining feature of the EU_HSHRs.
- Consistency with Image: The caption accurately identifies the object in the image as an EU_HSHR and provides additional information about its location (The British Museum, London) and museum number, which is consistent with the type of object depicted in the image.
- Overall Quality: The caption is informative and provides specific details about the object's museum location and identification number, but it lacks the integration of the highlighted segment, which affects the overall quality of the caption in the context of the task.

[ASSISTANT2-Reasoning]:
- Relevance with Context: Assistant 2's caption directly references the context by repeating the title "Euphrates Handmade Syrian Horses and Riders" and includes the term "zoomorphic clay figurines," which is central to the context.
- Relevance with Highlight: The caption successfully incorporates the highlighted segment "zoomorphic clay figurines," directly addressing the key feature of the EU_HSHRs.
- Consistency with Image: The caption is consistent with the image as it describes the object as a zoomorphic clay figurine, which is what is depicted.
- Overall Quality: The caption is coherent and relevant, effectively integrating the context and the highlighted segment with the image, resulting in a high-quality caption for the task.

[Comparison-Reasoning]:
When comparing the two captions, Assistant 2's caption is superior in all aspects. It is more relevant to the context as it includes the full title and the term "zoomorphic clay figurines," which is the highlighted segment. Assistant 1's caption, while providing specific museum information, fails to mention the highlighted segment, which is a critical part of the task. Assistant 2's caption is also consistent with the image, as it describes exactly what is shown. Overall, Assistant 2's caption is more coherent and integrates all required elements of the task more effectively than Assistant 1's caption.

[ASSISTANT1-Score]:
- Relevance with Context: 3
- Relevance with Highlight: 1
- Consistency with Image: 4
- Overall Quality: 2

[ASSISTANT2-Score]:
- Relevance with Context: 5
- Relevance with Highlight: 5
- Consistency with Image: 5
- Overall Quality: 5

Table 11: GPT-4 Prompt for the standard CIC Task

Prompt for CIC Task
CONTEXTUAL IMAGE CAPTIONING REQUEST

TASK: Based on the context and image description provided, please generate a contextual image caption that captures the essence of the image and its significance in the given context. For your reference, the Context Section will be given as the combination of a page title, a section title, and a section body. Some extra images other than the target image in the context section, if any, are included in the context section and described by [ImageCaptions] followed by text descriptions of the image to enrich the context. The image to be captioned itself is not directly accessible by the model; instead, a textual description of the image is given in the "Image Description" after the context section. Generate the proper caption for the image when the image is given together with the context. Please generate the caption directly without any explainations.

Context Section: [PageTitle] Hagerman horse [SectionTitle] Discovery [SectionText] A cattle rancher named Elmer Cook discovered some fossil bones on this land in Hagerman, Idaho. In 1928, he showed them to Dr. H. T. Stearns of the U.S. Geological Survey who then passed them on to Dr. James W. Gidley at the Smithsonian Institution. Identified as bones belonging to an extinct horse, the area where the fossils were discovered, called the Hagerman Horse Quarry, was excavated and three tons of specimens were sent back to the Smithsonian in Washington, D.C. Excavation of the fossils continued into the early 1930s. The Hagerman Horse Quarry floor grew to 5,000 square feet (460 m^2) with a backwall 45 feet (14 m) high. Ultimately five nearly complete skeletons, more than 100 skulls, and forty-eight lower jaws as well as numerous isolated bones were found. Some paleontologists believed that such a large amount of fossils found in one location was because of the quarry area being a watering hole at one point. The waterhole could have been where the bones of the Hagerman horses accumulated as injured, old, and ill animals, drawn to water, died there. Other paleontologists think that an entire herd of these animals drowned attempting to ford a flooded river and were swept away in the current and ended up buried in the soft sand at the bottom.

Image Description: a zebra with a white background

Contextual Caption (do not add quotations surrounding the response):

Response:
Zebra-like ancestor of the Hagerman horse, symbolizing its evolutionary connection

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)
2. 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)
3. 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)
4. 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)
5. 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)
6. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81 (2004)
7. 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)
8. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2019)
9. 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)
10. OpenAI: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
11. 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 Linguistic (ACL). pp. 311–318 (2002)
12. 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)
13. 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)
14. 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)
15. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575 (2015)