





VisChronos: Revolutionizing Image Captioning Through Real-Life Events

Phuc-Tan Nguyen^{*1,2} , Hieu Nguyen^{*1,2} , Minh-Triet Tran^{1,2} 
, and Trung-Nghia Le^{**1,2} 

¹ University of Science, VNU-HCM, Ho Chi Minh City, Vietnam

² Vietnam National University, Ho Chi Minh City, Vietnam

{21120028, 21120068}@student.hcmus.edu.vn, {tmtriet,
ltnghia}@fit.hcmus.edu.vn

Abstract. This paper aims to bridge the semantic gap between visual content and natural language understanding by leveraging historical events in the real world as a source of knowledge for caption generation. We propose VisChronos, a novel framework that utilizes large language models and dense captioning models to identify and describe real-life events from a single input image. Our framework can automatically generate detailed and context-aware event descriptions, enhancing the descriptive quality and contextual relevance of generated captions to address the limitations of traditional methods in capturing contextual narratives. Furthermore, we introduce a new dataset, Event-Cap (<https://zenodo.org/records/14004909>), specifically constructed using the proposed framework, designed to enhance the model’s ability to identify and understand complex events. The user study demonstrates the efficacy of our solution in generating accurate, coherent, and event-focused descriptions, paving the way for future research in event-centric image understanding.

Keywords: Event-Based Image Captioning · Contextual Captioning · Real-World Semantics · Event Extraction · VisChronos framework.

1 Introduction

Image captioning aims to generate descriptive captions for images. However, most current methods tend to produce captions with a limited understanding of the image, focusing primarily on identifying objects, actions, and basic physical attributes [8, 9, 14]. These approaches fall short in conveying deeper context, as they lack the ability to infer meaningful information about the events or interactions taking place in the image. This limitation becomes especially evident when the goal is to describe not just what is visible but also the underlying story or context associated with the image.

* Both authors contributed equally to this research.

** Corresponding author.



Caption by Grit:

- . The letters are brown colorful letters
- . Girl wearing with long brown hair
- . Letter a on a white waller on the wall
- . Scarf around woman's neck
- . Pink flower on black dress
- . The word before on the wall

Our caption:At a special screening of "The Lost City" at The Whitby Hotel in New York City on March 14, 2022, Sandra Bullock, the film's star, posed against a backdrop depicting the film's title in bold, energetic colors. The event served as a pre-release promotional push for the film, which was set to hit theaters on March 25th. Bullock, known for taking her roles seriously, expressed her desire to dedicate time to her children following this project, hinting at a potential break from acting. Her striking dress, featuring colorful abstract designs and sequins, aligned with the film's likely comedic and adventurous tone. The upscale venue of The Whitby Hotel suggested a sophisticated audience and elevated the event's significance, further building anticipation for "The Lost City".

Fig. 1: Comparison between general caption by Grit [14] and our event-based caption, highlighting additional details such as location, date, identity, and event purpose.

In many cases, the generated captions are too superficial to capture complex scenarios where additional information, such as who is involved, what is happening, where and when the event took place, and its significance, is critical. As a result, these methods are inadequate for providing rich, informative captions that align with more sophisticated user needs, such as understanding or retrieving images related to real-world events. To overcome this, a new approach is needed—one that integrates contextual details and event-related information to create comprehensive, narrative-driven captions that go beyond simple object recognition.

Our research introduces the task of Event-Enriched Image Captioning (EEIC), which aims to generate captions that provide richer, more comprehensive information about an image. This approach is demonstrated through a sample depicted in Fig. 1, where we showcase the result caption that our method generates. This example illustrates the enhanced descriptive quality and contextual depth that EEIC can bring to image captioning. These captions go beyond simple visual descriptions by offering deeper insights, including the names and attributes of objects, the timing, context, outcomes of events, and other crucial details—information that cannot be gleaned from merely observing the image. This approach facilitates the creation of more coherent and detailed narratives, capturing not only the visible elements but also the underlying context and significance of the scene, ultimately offering a more complete understanding of what the image represents.

The core idea of our approach is to harness event-related information from credible sources while leveraging the reasoning capabilities of both vision-language models and large language models (LLMs). We propose a fully automated four-step framework, called VisChronos, to analyze both the visual content and the

temporal, event-based aspects of the scene, ensuring a more comprehensive understanding. VisChronos operates through a systematic process designed to extract, analyze, and synthesize information from images and associated events. First, a vision-language model identifies and describes the most important aspects of the image, including both those specified by prompts and those deemed important by the model itself. Next, in the second step, an LLM generates questions about the image based on the key aspects identified in the first step, which includes both mandatory and optional questions. In the third step, another LLM answers these questions using event information that we provide. Finally, a separate LLM synthesizes and processes the information from the previous three steps to infer and generate the final caption for the image. Unlike traditional models that rely on learning from a massive dataset and risk generating information that may be fabricated or irrelevant, our method addresses this issue by incorporating factual sources: credible, human-authored articles that provide real, context-rich information. By drawing directly from authentic articles and aligning this information with image content, our framework ensures that captions accurately represent real events in human history. This framework is designed to establish an efficient information mining flow, systematically dividing different stages of information extraction across each step. This approach ensures that the framework can mine the most useful and relevant information at each stage, ultimately resulting in rich and contextually accurate captions.

Extensive human evaluations of captions generated by the VisChronos reveal that they are comparable to captions crafted by human annotators. These machine-generated captions were particularly praised for their completeness, coherence, conciseness, and the inclusion of relevant information not explicitly visible in the images.

Using VisChronos, we have created a dataset named EventCap, consisting of 3140 event-based image-caption pairs. Each pair has been carefully curated using images and related information sourced from a wide range of credible articles. This collection serves as a valuable resource for training and evaluating the performance of image captioning models in understanding and describing complex real-world events. To the best of our knowledge, no similar dataset exists that is specifically designed for the task of event-enriched image captioning, making EventCap a unique and essential resource for advancing research in this area.

Our main contributions can be summarized as follows:

- We introduce the new task of Event-Enriched Image Captioning (EEIC), which generates captions that provide profound insights into both the image and the events depicted, including information that cannot be inferred solely from the image.
- We propose a novel four-stage framework, namely VisChronos to solve the task of EEIC, combining vision-language models and LLMs to generate detailed, context-rich captions.
- We publicly release EventCap (<https://zenodo.org/records/14004909>), a dataset consisting of 3140 image-caption pairs, processed from 491 CNN articles.

2 Related Work

2.1 Dense captioning

Dense captioning, which aims to generate detailed descriptions for objects in scenes or videos, remains a challenging task. Over the years, several methods have been developed to address this problem, each employing distinct approaches and making notable contributions. Chen et al. [5] introduced Scan2Cap, an end-to-end method for dense captioning in RGB-D scans, utilizing 3D point cloud inputs to generate bounding boxes and corresponding object descriptions. Building on this, Wang et al. [13] presented PDVC, a framework that formulates dense video captioning as a task of set prediction, enabling efficient parallel decoding. Aafaq et al. [1] proposed the ViSE framework and VSJM-Net, which leverage early linguistic information fusion to model word-context distributional properties for improved dense video captioning. Shao et al. [10] further advanced the field by incorporating textual context-aware methods that generate diverse and context-rich captions. Similarly, Jiao et al. [7] presented MORE, a model that captures complex scene relations for superior captioning accuracy. Furthermore, Wu et al. [14] developed GRiT, a generative region-to-text transformer model that emphasizes object understanding in dense captioning tasks.

Despite the advancements made by these approaches, they typically generate conventional captions that lack real-world semantic depth, as they rely solely on information from the image or video. Moreover, most existing methods produce captions in a single pass through learned representations, which can result in missing critical details. In contrast, our method continuously supplements the captioning process through an interactive dialogue between models, allowing for the extraction of more nuanced and semantically rich information.

2.2 Large Language Models

Large Language Models (LLMs) have garnered significant attention due to their ability to generate human-like text and solve complex tasks across various domains. GPT (Generative Pre-trained Transformer), developed by OpenAI, is one of the most well-known LLMs. Floridi et al. [6] introduced GPT-3, a third-generation autoregressive language model that generates human-like text using deep learning techniques. They explored its nature, scope, limitations, and potential consequences, emphasizing that GPT-3 is not intended to pass complex mathematical, semantic, or ethical tests. Brown et al. [4] further demonstrated that scaling language models, such as GPT-3 with 175 billion parameters, significantly improves few-shot performance across various tasks, sometimes outperforming state-of-the-art (SOTA) approaches. In subsequent developments, Achiam et al. [2] presented GPT-4, a large-scale multimodal model capable of processing both image and text inputs to generate text outputs, marking a significant advancement in multimodal language modeling. Meanwhile, Yang et al. [15] explored GPT-4V, expanding GPT-4’s capabilities to include vision-based tasks, opening new avenues for large multimodal models. In addition, Wang et al. [12]

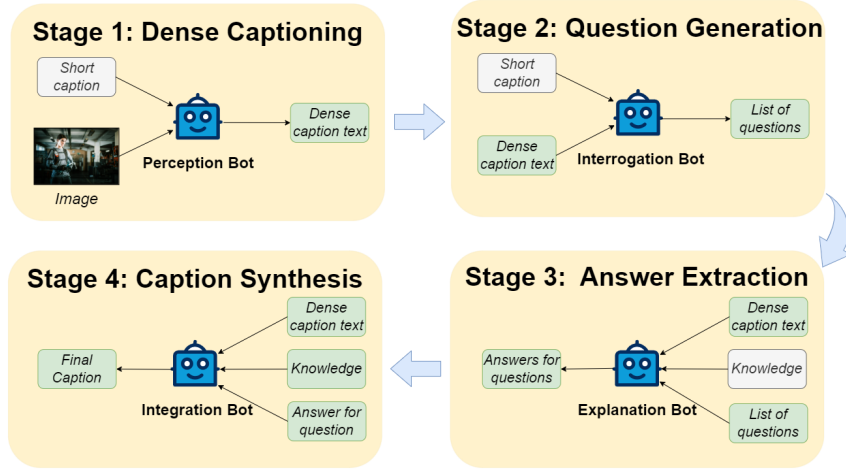


Fig. 2: VisChronos framework for Event-Enriched Image Captioning.

evaluated the trustworthiness of GPT models, concluding that GPT-4 is generally more reliable than GPT-3.5 but remains vulnerable to adversarial attacks, such as jailbreaking or misleading prompts.

Gemma is a lightweight, SOTA open model that builds upon the research and technology developed for Gemini models. As outlined in recent studies, Gemma has exhibited strong performance across various benchmarks for language understanding, reasoning, and safety [11]. Notably, Gemma has been rigorously evaluated alongside other large language models (LLMs) to assess its capabilities across different languages, modalities, models, and tasks [3]. The advancements in Gemma are a result of ongoing improvements in multimodal research, with enhancements in understanding and processing non-textual data inputs such as images and speech, which significantly contribute to its versatility in both academic and practical applications. Furthermore, by being available as an open-source model, Gemma provides accessible, cutting-edge AI technology to the broader research community, while also offering paid solutions for advanced functionalities and commercial deployment [11].

In this framework, we integrate the use of dense captioning models with both paid and open-source LLMs, including models such as GPT and Gemma, to leverage their combined strengths for enhanced performance across a range of tasks. By this framework, we also generate the EventCap dataset, the first dataset specifically designed for event captioning, providing a unique resource for accurately describing event contexts in diverse applications.

3 Proposed VisChronos framework

3.1 Overview

In this work, we introduce a novel multi-stage framework aimed at enhancing image captioning by generating contextually rich captions that go beyond describing the image. Our approach captures not only visual elements but also events, facts, and broader contextual information inferred from the image. As shown in Fig. 2, our framework consists of four stages, each of which is handled by a dedicated bot, designed to perform a specific task and feed information to the next stage. These stages are:

- **Stage 1 (Dense Captioning - Perception Bot)**: Generates a detailed description of the image, capturing objects, people, actions, and contextual elements.
- **Stage 2 (Question Generation - Interrogation Bot)**: Creates a set of questions based on the image description to explore the event’s context and details.
- **Stage 3 (Answer Extraction - Explanation Bot)**: Extracts answers to the questions from external knowledge sources, ensuring that answers are provided only when the model is confident.
- **Stage 4 (Caption Synthesis - Integration Bot)**: Synthesizes the dense caption, answers, and external knowledge into a comprehensive and contextually enriched final caption.

Each stage is designed to progressively enrich the information related to the image, ultimately producing a caption that is both visually descriptive and contextually informative.

3.2 Stage 1: Dense Captioning - Perception Bot

In this stage, the Perception Bot is responsible for generating a dense caption that thoroughly describes the image’s contents, including objects, people, actions, and background elements. This process is guided by a carefully crafted instruction provided to dense captioning model. The instruction ensures that the caption captures all significant visual elements, with attention to detail, relationships between objects, and overall context. Therefore, the dense caption can serve as the foundation for the entire framework, offering a detailed and structured description of the scene. This stage results in a rich description of the image with detailed information of objects, actions, and relationships, that serves as input for subsequent stages.

3.3 Stage 2: Question Generation - Interrogation Bot

The second stage is crucial for extracting the deeper context and events surrounding the image. This stage focuses on generating a set of structured questions that aim to explore various aspects of the event or scene depicted in the

Table 1: Key aspects covered by the questions in Question Generation stage.

Category	Description
Time	Questions addressing when the event occurred or its timeline to understand the temporal context.
Context or Reason	Questions exploring the circumstances or motivations behind the event, seeking to clarify why the event took place.
Main Events	Questions focusing on the central actions or occurrences within the event, ensuring a clear understanding of the primary narrative.
Outcome	Questions that inquire about the result or conclusion of the event, aiming to highlight the final impact or resolution.
Impact	Questions exploring the broader consequences or effects of the event on individuals, groups, or larger contexts.
Objects or People	Several questions delving into the key people or objects mentioned in the dense caption, ensuring that all significant elements are covered.
Special Figures	Specific questions about notable or important figures involved in the event, shedding light on their roles and influence.
Emotions and Reactions	Questions designed to explore the emotional states or reactions of the people in the image, providing insight into the human element of the scene.
Background Details	Questions addressing the setting or background elements in the image, helping to paint a fuller picture of the environment in which the event takes place.
Future Implications	Questions speculating on the potential future outcomes or ramifications of the event, aiming to place the event within a broader temporal and societal context.

image. The questions are designed to extract more information from the accompanying article or external sources, leading to a comprehensive understanding of the depicted event.

A structured approach is employed to generate questions that comprehensively address the event or context depicted in the image. The questions are crafted to ensure that all key aspects of the scenario are explored, providing a robust foundation for the subsequent explanation and synthesis stages as shown in Table 1.

The design of the question-generation model ensures that each of these dimensions is adequately explored, leading to a comprehensive inquiry into the event depicted in the image. This stage serves as the foundation for the next phase, where the generated questions are used to retrieve detailed answers from external knowledge sources.

3.4 Stage 3: Answer Extraction - Explanation Bot

The third stage focuses on extracting detailed answers to the questions generated in Stage 2 by leveraging external knowledge sources such as articles or accompanying text related to the image. The Explanation Bot ensures that the answers are accurate and directly relevant to the event depicted in the image.

This stage requires a highly precise approach to ensure that the model provides accurate answers grounded in the available information. A key principle guiding the model's behavior during this phase is certainty. The model is instructed to answer questions only when it is 100% confident in the information it provides. This ensures that the answers are factually reliable and directly linked to the knowledge from the article or dense caption.

Certainty Rule: If the model is unable to locate relevant information or if it is unsure about the correctness of an answer, it must respond with "no information." This instruction prevents the model from speculating or providing potentially misleading or inaccurate answers.

Information Retrieval: The model focuses on extracting factual data from the article or other external knowledge sources and cross-referencing this information with the dense caption to ensure consistency and accuracy. The goal is to answer each question as fully as possible, but only when the necessary information is available and can be confidently inferred.

This strict adherence to certainty ensures that the answers provided are reliable and fact-based, maintaining the integrity of the image-captioning process. By instructing the model to explicitly state "no information" when necessary, the framework avoids overgeneralization or the inclusion of speculative answers.

3.5 Stage 4: Caption Synthesis - Integration Bot

The final stage synthesizes the information gathered from previous stages into a comprehensive caption that reflects both the visual content of the image and the broader contextual information.

The instruction provided to the model for this stage ensures that it combines the dense caption, external knowledge, and question responses to generate a detailed and contextually enriched final caption. This synthesis produces a narrative that reflects both the visual and factual elements of the image. The final caption aims to describe the event depicted in the image, taking into account the knowledge gathered in the previous stages.

Output of this stage is a final caption that provides a rich, event-based description of the image, integrating visual details with contextual information from external sources.

3.6 Detailed Implementation

Our VisChronos framework is designed with flexibility, allowing integration with a variety of dense captioning models and LLMs, such as GRIT, GPT, Gemma, Bard, etc. This adaptability ensures that different models can be used

for various tasks within the framework. However, for the creation of the EventCap dataset, we specifically employed GPT-4o as the vision-language model for dense captioning in the first step, and utilized the Gemma-2-9b-it (open-source version) model for stages 2, 3, and 4 to handle question generation, answer extraction, and final caption synthesis.

4 Proposed EventCap Dataset

4.1 Dataset Construction

To build the EventCap dataset using our VisChronos framework, we required both images of real-world events and the corresponding event-related information, which we refer to as "knowledge." Specifically, we crawled data from 491 articles published by CNN between 2014 and 2022, covering a wide range of categories, including *business*, *entertainment*, *health*, *news*, *politics*, and *sport*. For each article, we collected the full textual content and all images included within the article.

The dataset creation process involved processing each image along with the corresponding article content, which served as external knowledge, through the VisChronos framework. The images were sourced from reputable news sites and captured by human photographers, ensuring authenticity and real-world relevance. The framework then generated captions for each image, using the article's content to ensure contextually relevant and detailed event descriptions. Captions generated by our method, as evaluated in Section 4.3, achieved a quality level comparable to human-written descriptions. This procedure formed the foundation for creating each image-caption pair in the dataset. Examples of image-caption pairs generated by our framework are illustrated in Fig. 3.

4.2 Dataset Specifications and Statistics

The EventCap dataset comprises 3140 image-caption pairs, with each pair considered a single sample, specifically curated to enhance the understanding of complex, event-driven content. The samples in the EventCap dataset are distributed across the years 2014 to 2022, covering a wide range of categories, including *business*, *entertainment*, *health*, *news*, *politics*, and *sport*; each categories having different sections. The images prominently feature current events, capturing significant moments in these years. These images vary widely in size, from 532 to 5440 pixels in width and 338 to 4618 pixels in height. Captions are generated using our VisChronos framework, designed to distill and articulate the essence and nuances of each event in a narrative style. These captions are extensive, ranging from 50 to 793 words with an average length of 106 words, and are structured in one or more paragraphs to convey detailed, context-rich descriptions of the depicted scenes. Fig. 4 presents sample distribution by years in the EventCap dataset while Fig. 5 breaks down the sample distribution by categories, illustrating the dataset's diversity.



The 2018 Girlboss Rally in Los Angeles, as captured in this image, was a significant event celebrating and empowering women in business, particularly within the male-dominated field of venture capital. The presence of Kara Nortman, Tracy Gray, and Nicole Quinn—all prominent venture capitalists—on stage underscores the growing influence of women in this sector. Their animated discussion, highlighted by Kara's gestures and Tracy's thoughtful expression, suggests a meaningful exchange of ideas and perspectives. This image encapsulates the event's spirit of collaboration, knowledge-sharing, and female empowerment within the venture capital industry.



The image captures the energy of an esports festival held in Abidjan, Ivory Coast, in 2018. The scene depicts young men intensely focused on competing in a popular soccer video game, likely FIFA. Their absorbed expressions and the setup—rows of monitors with gamers attentively controlling controllers—highlight the competitive nature of the event. This snapshot reflects the growing popularity of esports in Africa, as evidenced by the dedicated event space, the visible presence of spectators, and the diversity of gaming attire. This event is part of a larger trend, with Africa's gaming industry projected to grow significantly in the coming years, spearheaded by passionate gamers like Thabo "Yvng Savage" Moloi from South Africa, who recently became the first African to be sponsored by Red Bull. This festival in Abidjan symbolizes a passionate local community actively engaging with and shaping this exciting global phenomenon.



This image portrays David Minyatso, a widower, seated with his two daughters in their modest village home in Malawi. The warmth and intimacy of the scene offer a glimpse into the daily life of this family grappling with the profound loss of their matriarch, Selina, who died due to complications from an unsafe abortion. Minyatso's gaze and posture suggest resilience and a sense of responsibility in caring for his children. The children's innocent expressions highlight the vulnerability and innocence lost with Selina's untimely death, emphasizing the human cost of restrictive abortion laws and the desperate need for safe and legal access to reproductive healthcare in Malawi. This image serves as a poignant reminder of the far-reaching consequences of these laws, affecting not only the women who seek abortions but also their families and communities.



This image captures a heartwarming moment of national pride in Belgrade as fans enthusiastically welcome tennis superstar Novak Djokovic back to Serbia after his deportation from Australia. Djokovic, a celebrated figure in Serbian sports, was met with an outpouring of support, with fans waving Serbian flags, holding banners, and displaying emotional signs expressing their heartfelt admiration. The energy and enthusiasm of the crowd reflect the deep connection between Djokovic and the Serbian people, highlighting his status as a national hero. His return home emphasizes the importance of national identity and the unifying power of sports, particularly during challenging circumstances. This event showcases the profound impact of a successful athlete on their home country and the unwavering support they receive from their people.

Fig. 3: Examples of image-caption pairs generated by VisChronos.

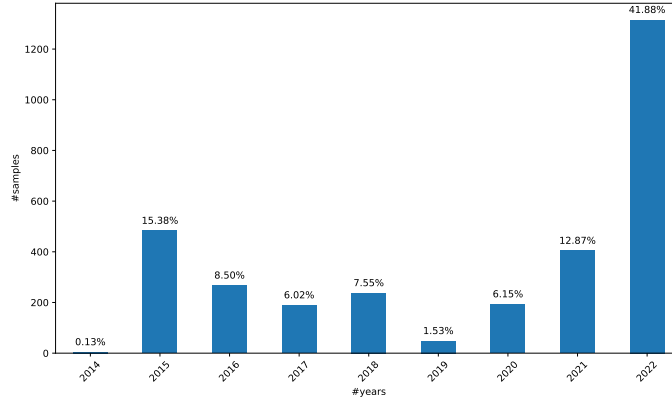


Fig. 4: Sample distribution in EventCap dataset by year (best view in color & zoom-in).

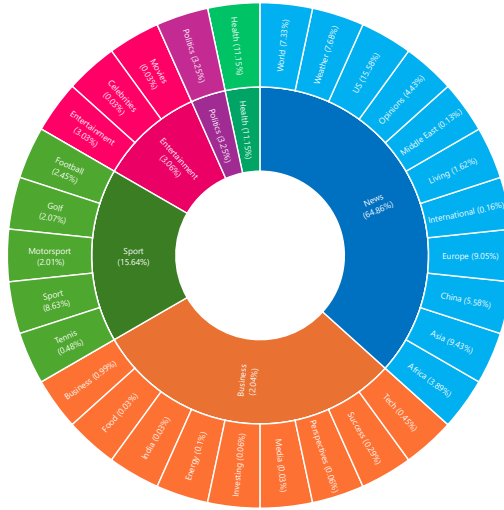


Fig. 5: Sample distribution in EventCap dataset by category and section (best view in color & zoom-in).

4.3 User Study

Since this task is novel, the existing metrics typically used to evaluate traditional models are not directly applicable. Furthermore, as this is the first dataset specifically created for this task, there is no comparable dataset available for direct benchmarking. To address this, we conducted a user study aimed at assessing both the effectiveness of our method and the quality of the generated dataset. Additionally, the study aimed to compare the quality of captions generated by

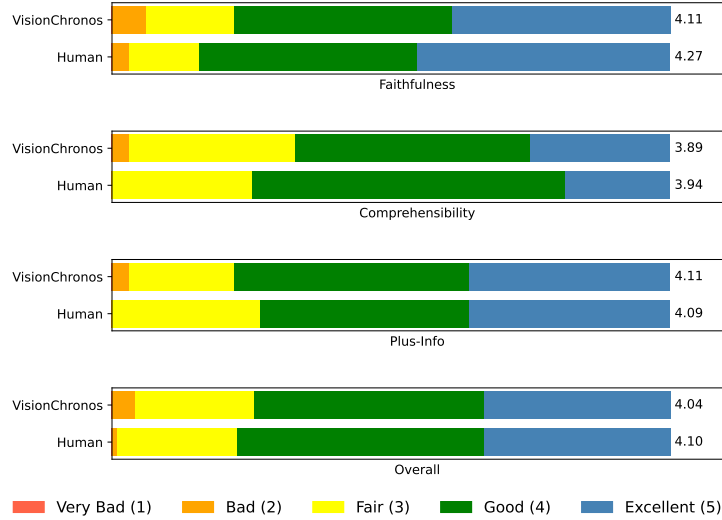


Fig. 6: Comparative performance of human and VisChronos in writing event-based image captions across metrics. Our proposed method can achieve human-level writing (best view in color & zoom-in).

our framework with those written by humans, highlighting the strengths and weaknesses of each approach.

Participants: We invited 10 participants (8 males, 2 females, aged 18 to 30) to participate in our study. Participants were selected from diverse backgrounds, including professionals in media, journalism, and computer science, to ensure a comprehensive and well-rounded evaluation.

Apparatus and procedure: Our study was conducted both online and on-site in our lab, where participants completed the tasks. Each participant received clear instructions on evaluating captions and writing their own for comparison. They were required to spend at least 4 minutes evaluating and 5 minutes writing the caption for each image including reading the corresponding article. The total time for the study sessions was approximately 120 minutes per participant.

First, participants were asked to write their own captions for 5-10 images from 2 articles. After that, participants evaluated the quality of captions for approximately 10-15 images from 4 different articles. Half of the captions were generated by our framework, while the other half were written by humans (i.e., other participants).

The participants were asked to rate the performance of the captions on a scale of 1 to 5 across three metrics, based on their individual perspectives. The comparison was based on several key metrics:

- **Faithfulness:** Whether the content of the caption fully describes the key events depicted in the image.
- **Comprehensibility:** Whether the caption is concise, easy to read, and free of unnecessary information.
- **Plus-Info:** Whether the caption provides additional useful information that cannot be inferred directly from the image.

Quantitative results We obtained quantitative results by averaging the ratings of each participant across three metrics and then averaging those results over all participants. These outcomes shown in Fig. 6 indicate that the performance of the VisChronos framework is not significantly different from that of human-written captions across evaluated metrics. This highlights potentials of our solution in real-life applications such as writing image description for newspapers, journals, books, etc. The user study results also demonstrates the high quality of our EventCap dataset in the development of event-based image captioning models.

Limitations Our method relies heavily on the quality of the accompanying information to produce accurate captions. Incomplete or poor-quality data can negatively impact the results. Additionally, the four-step process increases the time required, with an average of 1 minute needed to caption each image, which may limit scalability for large datasets.

5 Conclusion

In this paper, we introduced the VisChronos framework, a novel four-stage approach designed to tackle the new task of Event-Enriched Image Captioning (EEIC). This task aims to generate captions that not only describe the visual content of an image but also provide insights into the underlying events, incorporating information beyond what is directly observable. By leveraging a combination of vision-language models and large language models (LLMs), our framework iteratively refines the captioning process through interactive dialogue between models, resulting in semantically richer and contextually enhanced captions. Additionally, we presented EventCap dataset, consisting of 3140 image-caption pairs from 491 CNN articles, which serves as a strong foundation for future research in event-based image captioning and related tasks. To the best of our knowledge, EventCap is the first dataset specifically created for the task of event-enriched image captioning, establishing it as a pioneering resource and a strong foundation for future research in event-based image captioning and related tasks.

As part of ongoing efforts, we aim to significantly expand the EventCap dataset to 20,000 articles and 50,000 images, creating a comprehensive resource for the research community to advance event-enriched image captioning and multimodal learning. Additionally, we plan to refine the VisChronos framework into a multimedia tool that enhances human experiences by integrating visual,

textual, and contextual information. This evolution will support applications like event detection, real-time image analysis, and interactive storytelling, leveraging the framework’s ability to generate rich, event-driven captions.

Acknowledgement

This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

References

1. Aafaq, N., Mian, A., Akhtar, N., Liu, W., Shah, M.: Dense video captioning with early linguistic information fusion. *IEEE Transactions on Multimedia* (2022)
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., et al.: Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023)
3. Ahuja, S., Aggarwal, D., Gumma, V., et al.: Megaverse: Benchmarking large language models across languages, modalities, models and tasks. *arXiv preprint arXiv:2308.05698* (2023)
4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *NIPS* (2020)
5. Chen, D.Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2cap: Context-aware dense captioning in rgb-d scans. *arXiv preprint arXiv:2012.02202* (2020)
6. Floridi, L., Chiriatti, M.: Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines* (2020)
7. Jiao, Y., Chen, S., Jie, Z., Chen, J., Ma, L., Jiang, Y.G.: More: Multi-order relation mining for dense captioning in 3d scenes. In: *ECCV* (2022)
8. 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*. pp. 317–325 (2017)
9. Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., et al.: mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005* (2022)
10. Shao, Z., Han, J., Debattista, K., Pang, Y.: Textual context-aware dense captioning with diverse words. *IEEE Transactions on Multimedia* (2023)
11. Team, G., Mesnard, T., Hardin, C., et al.: Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2401.01234* (2024)
12. Wang, B., Chen, W., Pei, H., et al.: Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. *NIPS* (2023)
13. Wang, T., Zhang, R., Lu, Z., Zheng, F., Cheng, R., Luo, P.: End-to-end dense video captioning with parallel decoding. *arXiv preprint arXiv:2107.12589* (2021)
14. Wu, J., Wang, J., Yang, Z., Gan, Z., Liu, Z., Yuan, J., Wang, L.: Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2203.15806* (2022)
15. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v(ision). *arXiv preprint arXiv:2309.00332* (2023)