Are VLMs Ready for Critical Use Cases? Evaluating Zero-Shot Performance with Contextual Prompts

Oussama Zayene[1,*], Vincent Audergon[1], Jean Hennebert[1], Houda Chabbi[1] and Benoît de Raemy[2]

[1]*iCoSys, HEIA-FR, HES-SO University of Applied Sciences and Arts Western Switzerland*
[2]*Morphean SA*

Abstract

This study investigates Vision-Language Models (VLMs) for fire detection tasks, leveraging contextual prompts to assess their performance across various models. Notable results include, the Bunny model, which achieved a 76% F1-score, highlighting its effectiveness. These findings emphasize the impact of prompt engineering on performance while raising key questions about automating prompt optimization and selecting the most suitable VLMs based on task complexity, resource constraints, and real-world applicability.

Keywords

VLM, image captioning, visual question answering, contextual prompting

1. Introduction

The evolution of Generative AI, particularly through the use of transformers and self-attention mechanisms, coupled with advancements in hardware resources, has driven the development of Vision-Language Models (VLMs). These models integrate visual comprehension with natural language processing, enabling effective cross-modal understanding and interaction. Models like CLIP [1], InstructBLIP [2], Tag2Text [3] and Bunny [4] process images and videos alongside text, enhancing capabilities in tasks like image-text matching, and image/video captioning. These models excel in general tasks without domain-specific fine-tuning, driving interest in their use for real-world applications.

Yet, despite their impressive capabilities, VLMs are faced with significant challenges, particularly when applied to specialized, high-stakes domains like wildfire detection or environmental conservation. In these areas, errors can have severe consequences, as they require precise, context-aware judgments—something VLMs may struggle with due to their reliance on broad, unspecialized training data. Additionally, the cost and complexity of fine-tuning or retraining large-scale VLMs for niche applications are considerable, given that these models often contain billions of parameters and demand extensive computational resources. A key alternative consists of exploiting their zero-shot learning (ZSL) capability, where the model addresses new tasks without additional training. Instead of altering the model itself, conditioning its outputs through strategically designed input prompts can effectively guide its outputs, allowing for more domain-specific interpretations without the need for retraining.

A pressing question is can domain-specific tasks be effectively handled by leveraging prompt engineering, or are there inherent limitations to this approach? Another important consideration is how to develop robust evaluation methods for assessing VLM outputs in real-world scenarios where traditional reference-based metrics may not apply? To explore these questions, we established a concrete use case and selected a set of cutting-edge VLMs, each based on a different approach in the field. The use case is centered around specific contextual prompts (CPs) defined by users to reflect their interests in the images. Our experimental study is limited to images to streamline the evaluation process.

AI days HES-SO '25 January 27–29, 2025, Switzerland
*Corresponding author.
oussama.zayene@hefr.ch (O. Zayene); vincent.audergon@hefr.ch (V. Audergon); jean.hennebert@hefr.ch (J. Hennebert); houda.chabbi@hefr.ch (H. Chabbi)
https://www.heia-fr.ch/fr/recherche-appliquee/instituts/icosys/ (J. Hennebert)
0000-0001-9529-925X (O. Zayene); 0000-0002-5616-6830 (J. Hennebert); 0000-0001-7087-8108 (H. Chabbi)

The paper is organized as follows: Section 2 reviews the state of the art in VLMs, examining their strengths and limitations. Section 3 presents our approach, which aims to identify the most effective VLM for a given use case by evaluating each model's ability to correctly rank CPs based on their relevance to visual content. Section 4 details our experimental results and evaluation. Finally, Section 5 concludes with a summary of our findings and potential avenues for future research.

2. State of the Art in Vision-Language Models

Recent advancements in Generative AI have led to the emergence of various models that integrate visual content with natural language understanding. These models can be broadly categorized into three groups: Image-Text Matching, Multimodal-to-Text Generation and Chat-Centric models.

2.1. Image-Text Matching Models

This category includes VLMs [1, 5, 6, 7] focused on aligning visual content with textual descriptions through shared embedding spaces. Notably, CLIP (Contrastive Language-Image Pretraining) [1] was among the earliest models to leverage a dual-encoder architecture for this task, mapping images and text into a joint embedding space using contrastive learning. CLIP's strength lies in its ability to generalize to new tasks without task-specific fine-tuning, making it highly effective for zero-shot image classification, visual search, and image retrieval. Another interesting model in this category is ALIGN [5], which also employs a dual-encoder architecture by combining an EfficientNet-based [8] image encoder with a BERT-based [9] text encoder. This VLM leverages contrastive learning to align visual and textual embeddings, achieving good performance in tasks such as image-text similarity.

While having remarkable success, these models face several limitations. Among them, their reliance on contrastive learning may fail to capture deeper semantic relationships between images and text, leading to misalignment in complex scenarios.

2.2. Multimodal-to-Text Generation Models

The second category includes Multimodal-to-Text models [10, 11, 2, 3], which focus on producing textual descriptions, such as captions, from visual inputs. Unlike image-text matching models, which align visual data with text, these models are generative, meaning they create new text from visual inputs. A key example is Flamingo [11], a model designed for multimodal few-shot learning. Another interesting model is Tag2Text [3], which generates captions from images by leveraging auxiliary information such as tags or keywords. Models like BLIP (Bootstrapping Language-Image Pretraining) [12] and InstructBLIP [2] have advanced this field by incorporating instruction-based pretraining, enabling the models to handle tasks like captioning and visual question answering (VQA).

These models are particularly useful in applications like automated content generation and accessibility, where generating meaningful textual content based on images is essential. However, while they excel in generating grammatically correct text, the quality and specificity of the generated content can vary depending on the complexity of the visual input and the clarity of the prompt.

2.3. Chat-Centric multimodal Models

Chat-centric models are a class of VLMs designed for interactive multimodal dialogue systems. These models have seen rapid advancements, particularly with the rise of LLMs, enabling more sophisticated integration of V-L capabilities for dynamic, interactive dialogues. VisualGPT [13] was an early example, combining the power of GPT-3 [14] with vision transformers to enable natural, context-aware dialogue based on visual content. Following this, BLIP-2 [15] built on similar principles, enhancing multimodal dialogue with instruction-following capabilities, which improved its ability to provide more precise responses. LLaVA [16] integrates a vision encoder with a LLM backbone based on Vicuna, a high-performance variant of LLaMA2 [17], enabling more nuanced and interactive multimodal dialogues.

Bunny model [4] integrates LLaMA-3-8B [17] language model with SigLIP vision encoder [18], enabling multimodal capabilities for tasks like VQA and context-aware dialogue generation.

These models are particularly suited for more complex, conversational tasks, such as answering questions about images, and image-based reasoning. However, they face some limitations. Despite their conversational abilities, they often struggle with contextual accuracy and can generate irrelevant or imprecise responses when dealing with complex visual content.

3. Proposed Approach

In this section, we outline our approach for assessing the effectiveness of VLMs in domain-specific tasks using predefined CPs. The goal is to determine the best VLM for a given use case by evaluating its ability to align CPs with visual content. We focus on CLIP [1], Tag2Text [3], and Bunny [4], which belong to different categories of VLMs.

The use case we explore focuses on fire detection across diverse environments, including buildings, forests, and roadways. To this end, we define a set of CPs that capture various aspects of fire-related content, such as "a building on fire" and "flames coming from a vehicle". These prompts form the basis for evaluating the pipeline's ability to identify the most relevant CP by analyzing the image—either through CLIP's matching process or by generating captions with Bunny [4] or Tag2Text [3], and then comparing their alignment with the predefined CPs, as illustrated in Figure 1.

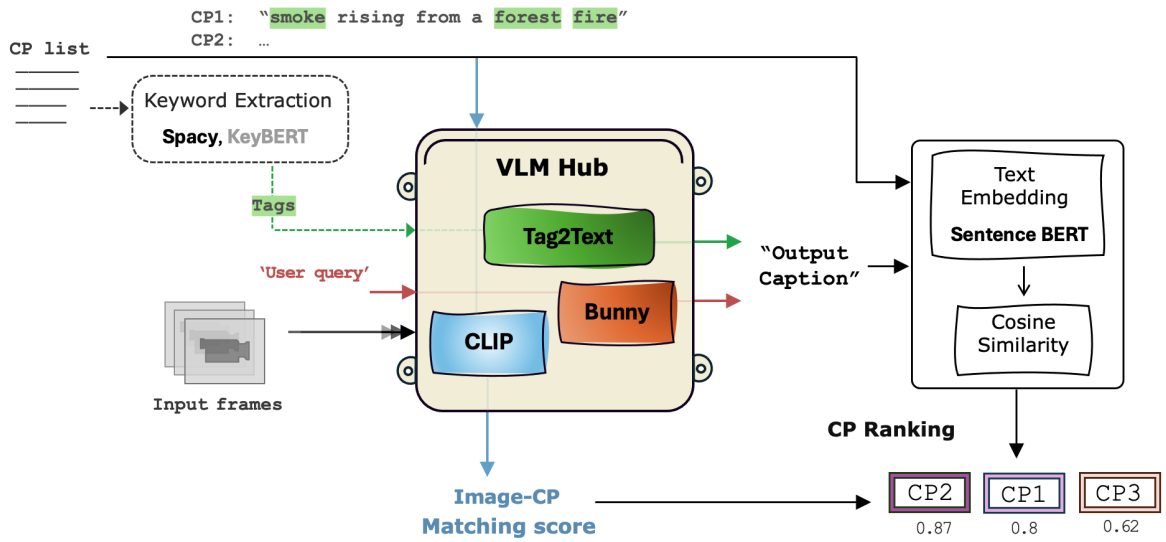

Figure 1: Proposed pipeline for image captioning and V-L matching using VLMs to classify input frames into predefined CPs. Blue pathways represent CLIP's input text (CP) and output score, while Green and Red indicate Tag2Text and Bunny pathways. Dashed lines show steps exclusive to Tag2Text's contextualized mode.

The core of our pipeline is the **VLM Hub**, which integrates multiple models. In its current implementation, the pipeline supports the manual selection of a single VLM at a time, based on the specific task at hand. Each model within the hub processes and analyzes the visual content in its own way:

- **CLIP** computes a matching score for each CP, identifying which corresponds the most closely to the image. A higher score indicates a stronger semantic match.
- **Tag2Text** operates in two modes. In its default mode, it generates descriptive captions for images without any additional context. However, in its contextualized mode, the model leverages user-specified tags derived from CPs via SpaCy method [19] (highlighted by dashed lines in Figure 1) to guide the captioning process towards producing context-aware descriptions.
- **Bunny** is designed for image-based dialogue, where it generates captions in response to user query "Describe the image concisely". Unlike other models, it doesn't generate captions automatically but instead relies on the user's input to guide the description.

The last stage of our pipeline involves converting both the captions generated by Tag2Text [3] or Bunny [4] and the original CPs into embeddings using Sentence-BERT [20]. These embeddings are compared using cosine similarity to evaluate how closely each CP aligns with the generated captions. CPs are then ranked based on their similarity scores, with a predefined threshold of 0.5 determining which CPs are considered the most relevant to the image description.

4. Experimental Results

In this section, we present the experiments conducted to assess the performance of VLMs. We investigate the interaction between predefined CPs and each VLM's generated captions, as an alternative evaluation method due to the absence of GT captions. This problem can thus be reframed as a *classification task* to predict most relevant CP from a predefined set, such as **vehicle burning (CP1)**, **building burning (CP2)**, **forest burning (CP3)**, and a **neutral category (#)**. After experimenting with different CP sets, the final selection was made through trial-and-error prompt engineering for the evaluation phase.

Figure 2: Examples from the created dataset depicting different fire (and neutral) scenarios.

4.1. Dataset Description

A dataset of 205 fire-related images has been compiled for testing, sourced from publicly available datasets such as OnFire [1] and ForestFire [2], as well as public web sources. It includes 57 neutral images. Figure 2 shows examples from this dataset, which covers various fire scenarios in different environments (wild and urban), lighting conditions (day and night) and depth variations. This diversity ensures that the dataset is designed to support a comprehensive evaluation of the models across different scenarios.

The annotation process links each image to the corresponding CP. Each entry in the JSON GT file includes the image filename, the CP key, and optionally an object name (e.g., "Fire").

4.2. Experiments and Discussions

In this section, we present the quantitative results of our experiments, comparing the performance of five systems: a dummy model, CLIP, Tag2Text (in both default and contextualized modes), and Bunny. The evaluation metrics include Precision, Recall, and F1-score, calculated across the 4 predefined classes (CPs). The results, summarized in Table 1, reveal clear differences in performance among the VLMs.

[1] https://mivia.unisa.it/onfire2023/index.html
[2] https://github.com/EdoWhite/ViT-Forest-Fire-Detection/blob/main/RawFireData

Table 1
Performance comparison of different VLMs in terms of Precision, Recall, and F1-score.

VLM	Precision	Recall	F1-score
Dummy model	0.27	0.25	0.24
CLIP	0.54	0.60	0.53
Tag2Text (Default)	**0.79**	0.73	0.72
Tag2Text (Contextualized)	0.70	0.66	0.65
Bunny	**0.79**	0.77	**0.76**

Dummy model, with random confidence scores and fixed caption, serves as a baseline with expectedly low scores across all metrics. **CLIP** shows notable improvement, achieving a gain of 0.29 in F1-score. However, its reliance on direct matching between textual and visual embeddings may limit its ability to handle fine-grained, domain-specific prompts. The **default Tag2Text** performs well with an F1-score of 0.72, demonstrating its capability to generate descriptive captions that align with predefined CPs. Interestingly, the **contextualized Tag2Text**, which uses CP-derived tags, underperforms by 7%. This indicates that while contextualization can enhance relevance in some cases (left part of Figure 3), it may also introduce noise or misalignment. Overall, this VLM still successfully detects critical elements such as the presence of fire, which is a key capability for applications like fire monitoring and emergency response, even if its contextual descriptions are less precise compared to Bunny. On the other hand, **Bunny** achieves the highest performance, with an F1-score of 0.76, demonstrating the strength of its cross-modality projection mechanism. This is supported by the qualitative results shown in Figure 3.

Figure 3: Examples of generated captions

Our experiments show that caption quality significantly impacts CP classification accuracy. However, there is a trade-off between accuracy and computational efficiency. Advanced VLMs like Bunny offer high accuracy but have high GPU demands, making them less practical for simpler tasks. In contrast, CLIP provides adequate performance with lower resource needs, while Tag2Text strikes a balance, achieving good results without the computational costs of more complex models.

5. Conclusions

This study highlights the potential of VLMs in domain-specific tasks, particularly fire incident classification using CPs. While no model excels across all metrics, Bunny effectively combines visual and textual data, whereas CLIP offers a computationally efficient alternative. This raises the question of how to determine the most suitable VLM —or combination of VLMs— based on the task complexity, and available computational resources? Evaluating the relevance of VLM-generated captions remains challenging without GT data, making traditional metrics unsuitable. Could human-in-the-loop strategies be necessary to maintain accuracy in critical applications?

Future work will focus on assessing caption relevance and refining model selection for specific tasks.

Acknowledgments

This work is part of the VideoCognition Innosuisse project, a collaboration between iCoSyS Lab. Morphean SA. Special thanks to the team members for their dedication and expertise.

References

[1] A. Radford, et al., Learning transferable visual models from natural language supervision, in: Proceedings of the International Conference on Machine Learning (ICML), 2021.

[2] W. Dai, et al., Instructblip: Towards general-purpose vision-language models with instruction tuning, in: Advances in Neural Information Processing Systems (NeurIPS), 2023.

[3] Z. Huang, et al., Tag2text: Guiding vision-language model via image tagging (2023).

[4] M. He, Y. Liu, B. Wu, J. Yuan, Y. Wang, T. Huang, B. Zhao, Efficient multimodal learning from data-centric perspective (2024).

[5] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: International conference on machine learning, PMLR, 2021, pp. 4904–4916.

[6] L. Yuan, S. Chen, Y. Zhu, M. Zeng, X. Liu, Q. Ma, X. Zheng, L. Li, D. Ramanan, L. Zhang, et al., Florence: A new foundation model for computer vision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[7] X. Zhai, J. Puigcerver, A. Kolesnikov, N. Houlsby, C. Riquelme, Lit: Zero-shot transfer with locked-image text tuning, arXiv:2111.07991 (2022).

[8] M. Tan, Q. V. Le, Efficientnet: Rethinking model scaling for convolutional neural networks, in: Proceedings of the International Conference on Machine Learning (ICML), 2019.

[9] J. Devlin, et al., Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019.

[10] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., Oscar: Object-semantics aligned pre-training for vision-language tasks, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK,, Springer, 2020, pp. 121–137.

[11] J.-B. Alayrac, et al., Flamingo: A visual language model for few-shot learning, arXiv (2022).

[12] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, 2022.

[13] J. Chen, H. Guo, K. Yi, B. Li, M. Elhoseiny, Visualgpt: Data-efficient adaptation of pretrained language models for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18030–18040.

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, E. Shinn, N. Shazeer, N. Shinn, et al., Language models are few-shot learners, arXiv:2005.14165 (2020).

[15] J. Li, et al., Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International Conference on Machine Learning, 2023, pp. 19730–19742.

[16] H. Liu, et al., Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[17] H. Touvron, G. Lample, S. Ruder, et al., Llama: Open and efficient foundation language models, arXiv:2302.13971 (2023).

[18] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, arXiv:2303.15343 (2023).

[19] M. Honnibal, I. Montani, S. V. Landeghem, et al., spacy: Industrial-strength natural language processing in python, 2020. URL: https://spacy.io/.

[20] N. Reimers, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv (2019).