

ComiCap: A VLMs pipeline for dense captioning of Comic Panels

Emanuele Vivoli^{1,2}, Niccolò Biondi², Marco Bertini², and
Dimosthenis Karatzas¹

¹ Computer Vision Center, UAB, Spain

² MICC, University of Florence, Italy

{evivoli,dimos}@cvc.uab.es

{name.surname}@unifi.it

Abstract. The comic domain is rapidly advancing with the development of single- and multi-page analysis and synthesis models. Recent benchmarks and datasets have been introduced to support and assess models' capabilities in tasks such as detection (panels, characters, text), linking (character re-identification and speaker identification), and analysis of comic elements (e.g., dialog transcription). However, to provide a comprehensive understanding of the storyline, a model must not only extract elements but also understand their relationships and generate highly informative captions. In this work, we propose a pipeline that leverages Vision-Language Models (VLMs) to obtain dense, grounded captions. To construct our pipeline, we introduce an attribute-retaining metric that assesses whether all important attributes are identified in the caption. Additionally, we created a densely annotated test set to fairly evaluate open-source VLMs and select the best captioning model according to our metric. Our pipeline generates dense captions with bounding boxes that are quantitatively and qualitatively superior to those produced by specifically trained models, without requiring any additional training. Using this pipeline, we annotated over 2 million panels across 13,000 books, which will be available on the project page <https://github.com/emanuelevivoli/ComiCap>.

Keywords: Dense Caption · Comics panels · Vision Language Models

1 Introduction

Comics represent a highly complex medium for computational analysis, yet they are easily understood by humans—except for individuals within the Blind or Low Vision community. Recent studies [15, 19, 23] have addressed this gap by developing dialog generation tasks to assist People with Visual Impairments (PVI). These tasks aim to transcribe all spoken text, sorted by appearance, and associate it with the corresponding character's name. To support these efforts and facilitate new methods, several benchmarks have been introduced [28, 29]

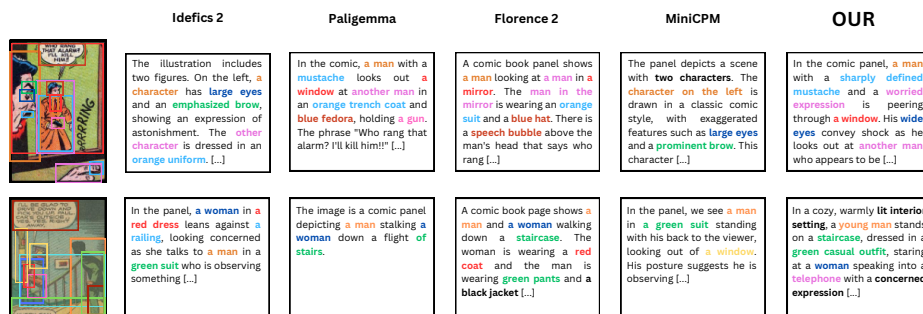


Fig. 1: VLMs dense captions compared to our pipeline.

that handle multiple tasks across various comic styles, from detection to dialog generation. However, when consuming a comic book solely through transcribed dialog, a crucial element is missing: context. Context, defined as a description of the scene’s happenings, can be extrapolated using recent Vision-Language Models (VLMs) through captioning [21]. A comprehensive model to assist PVI should generate comic dialog transcriptions while providing dense captions for specific panels and descriptions of characters when necessary.

Despite rapid advancements in dialog transcription [22, 23, 28], no research has adequately addressed the task of panel and character description, particularly through a grounded approach that includes bounding boxes associated with generated captions to enhance explainability.

In this paper, we focus on generating dense captions for comic panels using existing VLMs without additional training. Our contributions are as follows:

- We propose a two-stage metric based on automatic key-element extraction and BERT-score evaluation to assess the presence of important attributes in VLM-provided captions.
- We annotate 1.5k panels with captions and attribute lists, benchmarking existing open-source VLMs for comic panel caption generation using the proposed metric.
- We identify the top-performing Vision-Language model and propose a multi-stage pipeline that refines panel captions from fine- to coarse-grained, demonstrating superior dense captions compared to specifically trained models.

With the proposed pipeline, we annotate more than 2 million panels, providing dense captions to the research community.

The rest of the paper is structured as follows: in Section 2, we provide an overview of existing tasks in the comic domain, as well as an overview of visual language models. In Section 3, we detail our attribute extraction process starting from VLM captions. In Section 4, we justify and describe the proposed metric. Section 5 evaluates VLMs against the benchmark dataset using our metric and iteratively designs our pipeline. Finally, in Section 6, we present qualitative results and discuss future work that our approach enables.

2 Related Work

Comics tasks. In the field of comics, common tasks span from detection, segmentation, and element linking (such as speaker identification) to clustering (character re-identification). Recently, some works have focused on designing dialog generation tasks as a proof of concept for single-page comics analysis, starting from simpler atomic tasks [23] and later also extending to include names [22]. However, many of these tasks lack proper evaluation, leading recent works to provide metrics and benchmarks for detection, linking, and dialog generation tasks [28, 29]. Despite these advances, a story cannot be fully understood, especially by those who cannot easily access comic images, when relying solely on dialog generation and character naming. Numerous events occur within a scene and across the gutters [8]. A natural progression is the detailed panel description through captioning to provide context for every significant scene [21]. Recent VLMs have demonstrated surprising performance in generating both short and long captions, particularly in out-of-domain settings.

Vision-Language models. Recently, vision and language models have garnered significant attention. A common practice involves integrating a vision encoder into large language models (LLMs), employing various approaches [1, 16, 17, 33], effectively giving LLMs “eyes.” This simple approach can be extended with multiple modifications to accept interleaved (image-text) data [14], high-resolution images [16, 31], and generate text output [17, 18, 33], or even bounding boxes [4, 30] and segments, using VQ-GAN mask quantized vector indices [4] or actual segment perimeter indices [30]. These models are typically trained for a broad range of tasks, following (i) pretraining (only training the mapping operation from vision space to language space), (ii) finetuning (unfreezing the language decoder and, in some cases, also the vision encoder), and (iii) instruct fine-tuning/downstream task setting (unfreezing everything, or training with LoRA). This method slowly adapts the model to align image and text representations in the same space, enabling it to process and understand images and, in some cases, provide detection boxes that validate the model’s understanding, thereby aiding human reasoning about the detected elements.

Dense captioning. Recent advancements in image captioning, such as those by Vinyals et al. [27] and Anderson et al. [2], have laid the foundation for dense captioning, an advanced extension of traditional image captioning. Dense captioning [10, 11] involves generating detailed descriptions for multiple regions within an image, integrating object detection with captioning [7, 20, 24]. This method significantly enhances visual content description over conventional single-sentence captions. Dense captioning has evolved to include video applications [12] and has been incorporated into vision-language multi-task models like Florence2 [30], which are trained on extensive datasets such as FLR-5B. Despite these advancements, dense captioning remains challenging in certain media like comics, which are underrepresented in training datasets for vision-language models (VLMs).

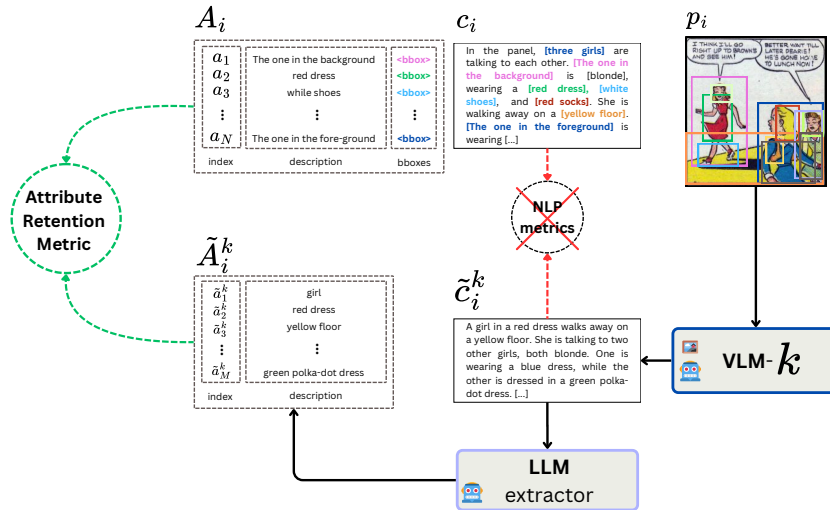


Fig. 2: Attribute extraction from VLMs captions.

3 Captioning and Attribute Extraction

Our goal is to generate a dense caption from a comic book panel that includes all important attributes related to the scene and characters, along with their bounding boxes. In this section, we describe how we utilize VLMs in a zero-shot manner to generate a list of attributes that can later be grounded (see Figure 2). These attributes are used to evaluate the caption’s quality against a designed test set of panel captions and attributes.

Captioning models. Among the recent VL models, the majority have been trained for captioning [5].

Idefics2 [14], a fully autoregressive architecture with 8 billion parameters, leverages the SigLIP-SO400M vision encoder [32] for robust visual processing and the Mistral-7B language model [9] for advanced language understanding capabilities. The model is trained using the OBELICS dataset [13] for foundational multimodal understanding and is further refined via instruction tuning with The Cauldron dataset, which includes a wide range of vision-language tasks.

The MiniCPM-llama3-V-2.5 model [18] integrates a sophisticated architecture blending elements from SigLIP-400M [32] and Llama3-8B-Instruct [6]. It was trained using a unique set of training data that includes CommonCrawl and Code Pretrain datasets, C4, and smaller contributions from sources such as Arxiv and Open Web Math, among others. For handling high-resolution images, the model employs a resampling technique that adjusts the input images to optimal sizes and resolutions without extensive padding or reshaping, maintaining efficiency and minimizing data distortion. Fine-tuning involves reinforcement learning from AI feedback (RLAIF-V), using AI-generated responses to refine

the model’s alignment with human-like reasoning, substantially reducing errors such as hallucinations. This model supports a wide array of tasks, including high-quality OCR for large-scale images and complex instruction following.

Grounding models. Some vision-language models also support generating bounding boxes associated with textual hooks.

PaliGemma [4] is a versatile 3B Vision-Language Model (VLM) that combines the capabilities of a 400M SigLIP vision encoder [32] with a 2B Gemma language model [25]. PaliGemma follows a prefix-LM masking paradigm, always prefixing image tokens followed by the prompts/questions, applying usual causal masking to the generated text. The model is designed to deliver bounding boxes and segmentation masks using two additional vocabularies: a 1024-bin vocabulary for x-y detection boxes and a 128 VQ-VAE tokenized single object mask tokens. The training involves multiple stages, starting with individual pre-training of unimodal components, followed by multimodal pretraining, resolution increase, and finally, task-specific transfer training, including captioning.

Florence2 [30] is a foundation model that adopts a unified, prompt-based approach to accommodate a diverse set of vision and vision-language tasks. It integrates a vision encoder and text-location tokenizers, which are provided as input to an encoder-decoder transformer in a sequence-to-sequence framework. The model is trained on the FLD-5B dataset, a meticulously curated collection featuring over 5.4 billion annotations across 126 million images. Florence2 processes multimodal inputs and outputs text and location tokens, enabling it to perform tasks ranging from object detection and image captioning to visual grounding and segmentation. Florence2 performs dense captioning in a two-step process: first, generating a caption of varying lengths, and second, using the caption for text-phrase grounding.

These models were selected for their variability in architecture, training procedure, fine-tuning datasets, and supported tasks. In particular, all these models support image captioning. As illustrated in Figure 2, given a panel $p_i \in P$, with P being the set of panel images, the first phase consists of generating a caption \tilde{c}_i for the panel p_i with the above models $x^k \in X$, with k being in $\{1, 2, 3, 4\}$. All models are prompted with the text “describe the image in detail.” These provided captions \tilde{c}_i^k , with longer or shorter text depending on the model’s specifics, are used to extract the attribute list.

Attribute extraction. The generated caption should retain all the panel’s objects and attributes (e.g., objects in the scene, important elements, day or night attributes) and characters’ objects and attributes (e.g., clothing, facial and body characteristics, expressions, specific poses) for every significant character in the scene. These are the objects and attributes elements we are interested in.

In the Natural Language Processing field, when comparing sentences to each other (in our case, raw captions), it has been observed that the METEOR metric most closely aligns with human preferences for multiple-choice captions given an image [3]. However, this can be misleading, as our preference is for captions that retain all important elements and attributes of an image, even if this sacrifices

some fluency and brevity. Thus, we propose a custom procedure that employs LLMs to perform the task of attribute extraction. Specifically, given the panel image p_i and the caption \tilde{c}_i^k generated by the model x^k , we aim to extract from \tilde{c}_i^k a set of attributes $\tilde{A}_i^k = \{\tilde{a}_1^k, \tilde{a}_2^k, \dots, \tilde{a}_M^k\}$ that is most similar to the ground truth attribute set $A_i = \{a_1, a_2, \dots, a_N\}$. The set of ground truth attributes has been used to generate the ground truth caption c_i . As described above, our goal is to obtain similar attribute sets and not necessarily BLEU, ROUGE, or METEOR-optimized captions. Thus, we analyse the attributes set \tilde{A}_i^k for all models k , that we obtain employing an LLM specifically prompted for performing this task. The prompt is provided in the project repository.

4 Attribute Retaining Metric

As illustrated in Figure 2, we are interested in a metric that can compare two sets of elements: one produced by the chosen model and the other being the ground truth. The metric should be able to associate the elements $a_i, i \in \{1, \dots, N\}$ with the predicted $\tilde{a}_j^k, j \in \{1, \dots, M\}$, even if they are not exact string matches. To achieve this, we design a two-step procedure. First, we compute a pairwise BERT-score among all possible (i, j) pairs. Since the BERT-score is bounded between 0 and 1, we choose a threshold to avoid misleading associations. The threshold τ is chosen among 0.5 and 0.99, for all models. All the detected associations are then substituted in the predicted attributes set by the respective ground truth elements, resulting in the modified set \tilde{A}_i^{*k} . The second step of the metric involves applying the Jaccard similarity, or intersection-over-union metric, to the sets of elements.

The complete pseudo code is provided in Algorithm 1.

Finally, we compute the predicted set accuracy ARM as the average Jaccard similarity: $\frac{1}{N} \sum_{i=1}^N J_i$.

5 Pipeline

Ablation. We have a collection of VLMs capable of captioning comic book panels. These model-generated captions are synthesized by an LLM, specifically GPT4o-mini, which extracts the attributes from the panel captions and provides a list of objects and attributes present in the captions. The prompt used to extract the attribute set is provided in the project repository. Finally, we obtain a Jaccard similarity score from the BERT-score filtered attribute sets, which we use to assess the model’s precision in retaining all important elements in the caption. The results are provided in Table 1.

As seen from the table, MiniCPM retains the highest ARM score, which is not in line with other metrics results such as METEOR, BLEU and RIUGE. This is thanks to the design of our metric, which emphasises the presence of attributes and objects, and not bigram exact co-occurrence. An additional post-processing technique applied to all models captions involves removing OCR-detected text from the caption. MiniCPM has been pre-trained for OCR transcription, which

Algorithm 1 Attributes Retaining Metric Calculation

```

1: procedure CALCULATEARM( $C, \tilde{C}^k, \tau$ )
2:   Input: Ground truth captions  $C = \{c_i | i = 1, \dots, |P|\}$ , predicted captions  $\tilde{C}^k = \{\tilde{c}_i^k | i = 1, \dots, |P|\}$  obtained with model  $x^k$ , vision language model  $x^k$ , threshold  $\tau$ 
3:   Output: Jaccard similarities  $J^k = \{J_i^k | i = 1, \dots, |P|\}$  with  $p_i$  being the panel and  $|P|$  being the size of panels set.
4:   for each  $i$  from 1 to  $|P|$  do
5:     Extract predicted entity sets  $\tilde{A}_i$  from predicted caption  $\tilde{c}_i$  using model  $x^k$ 
6:     Calculate BERT-score  $BS(\tilde{A}_i, A_i)$  for each  $A_i$ 
7:   end for
8:   for each  $i$  from 1 to  $|P|$  do
9:     Initialize cleaned set  $\tilde{A}^*_i \leftarrow \emptyset$ 
10:    Initialize cleaned set  $J_i \leftarrow \emptyset$ 
11:    for each element  $\tilde{a}$  in  $\tilde{A}_i$  do
12:      if  $BS(\tilde{a}, A_i) \geq \tau$  then
13:        Replace  $\tilde{a}$  with the matching element  $a \in A_i$ 
14:        Add  $a$  to  $\tilde{A}^*_i$ 
15:      else
16:        Add  $\tilde{a}$  to  $\tilde{A}^*_i$ 
17:      end if
18:    end for
19:    Calculate the Jaccard similarity  $Jacc(\tilde{A}^*_i, A_i)$ 
20:    Save it in  $J_i$ 
21:  end for
22:  Return:  $J_i$  for  $i \in 1, \dots, |P|$ .
23: end procedure

```

Table 1: Performance Metrics for Different Models

Model	ROUGE	BLEU	METEOR	ARM (ours)
PaliGemma	0.13 ± 0.02	$0.01 \pm <0.001$	0.05 ± 0.001	0.22 ± 0.11
Idefics2	0.19 ± 0.07	0.29 ± 0.13	0.19 ± 0.08	0.23 ± 0.10
Florence2	0.31 ± 0.12	0.17 ± 0.11	0.17 ± 0.09	0.24 ± 0.11
MiniCPM	0.38 ± 0.12	0.34 ± 0.07	0.29 ± 0.12	0.36 ± 0.11

is evident from its captions, thus this techniques might have impacted the results too.

Once the best VLM is chosen, we can explore an additional setting: improving the captions by separating panel and character captions. Specifically, we employ the DASS model [26] for detecting character boxes and provide the model with the panel and each of the character cut-out boxes. The MiniCPM model, as illustrated in the model description, has been trained for instruction following and refined with reinforcement learning techniques, making it suitable for following instructions. We leverage this property by providing two different prompts to the model: one for panels and one for characters. Both prompts are structured so that the output includes the caption and the attribute list together. We successfully parse the output into a CSV file and a text caption file using the appropriate prompt. The prompts are provided in the project repository.

Figure 3 provides an overview of the pipeline, differentiating the two prompts for panels and characters, and obtaining the captions and attribute lists simultaneously. Notably, this instruction-following property is not available in other VLMs. We empirically discovered that using the panel-characters setting is more favorable compared to using a single caption for the panel.

Text Grounding. To achieve dense captioning, we must address the inclusion of bounding boxes. Regardless of the chosen model, an additional VL model can be employed to detect each object in the attributes set. In the case of Florence2, the grounding model is Florence2 itself. However, for MiniCPM, we use GroundingDINO, which has demonstrated significant performance in zero-shot out-of-vocabulary detection, including in contexts like manga [23].

We apply GroundingDINO on the attribute list extracted by MiniCPM. As GroundingDINO sometimes misses objects due to them being unknown, we also ask MiniCPM to provide additional synonym terms to support the attributes in the attribute set.

Figure 3 provides an overview of the full pipeline, utilizing MiniCPM with the panel-characters setting and the final version of the prompts. In Figure 4, we provide examples of MiniCPM outputs, which we parse by searching for the “csv” and “caption” tags.

Examples of GroundingDINO-detected attributes are provided in Figure 5.

ComiCap Dataset. Copyrights present a significant issue in comics. However, some websites provide sufficiently old, out-of-copyright comic books that are

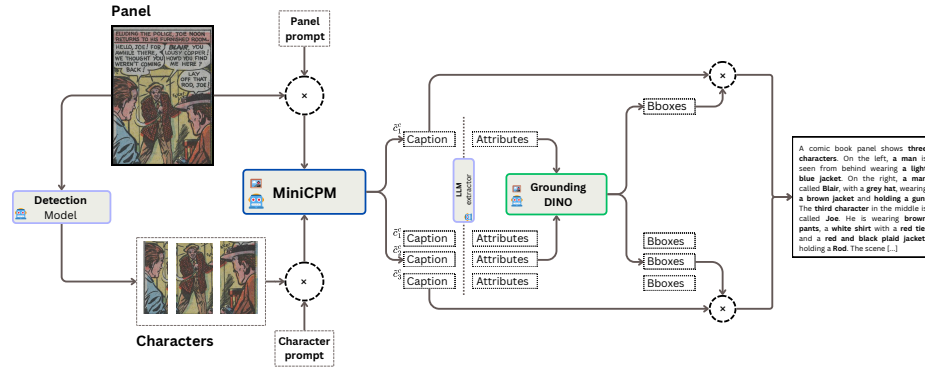


Fig. 3: Pipeline details for the MiniCPM model using the panel-characters setting.



Fig. 4: Examples of the MiniCPM output, ready to be parsed into attribute lists (csv) and captions (txt).

freely accessible and downloadable. We have collected over 13k books from the Digital Comic Museum website, which has been active since 2011 and contains a collection of more than 22k books. Some of these books were unavailable or corrupted, resulting in a final collection of 13k books. Recent work has trained (under limited training sets) and tested various comic style object detection models [29]. The authors demonstrated that, among single- and double-step CNN detectors and transformer models, the best-performing model for comic style is FasterRCNN (for panels) and DASS (for characters). Therefore, we adopted these models to automatically extract panels and characters in the DCM-13k dataset, similar to [8]. This procedure resulted in over 1.5M panels and 2.06M characters detected, culminating in a densely captioned comics panels dataset, which we provide to the research community. Some qualitative results of the generated captions are shown in Figure 6.

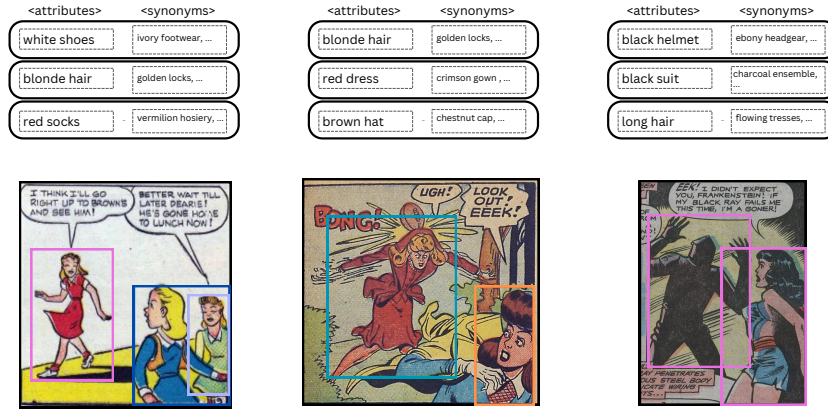


Fig. 5: Examples of GroundingDINO detections: correctly identified elements (left), missed ones (center), and wrongly detected objects (right).

6 Conclusion and Future Work

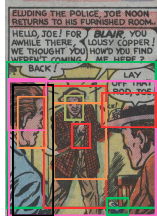
In this paper, we have demonstrated an automatic method for generating dense captions of comic panels without requiring additional training, leveraging existing Vision-Language Models (VLMs). We identified limitations in existing captioning metrics and proposed a novel attribute-based metric that combines BERT-score and Jaccard similarity to assess attribute coverage in captions. We benchmarked various VLMs using this metric on a specifically annotated test set. The best-performing model was further explored through prompt engineering to develop a comprehensive dense captioning pipeline, which also incorporates GroundingDINO for text grounding. To the best of our knowledge, this is the first work to generate dense captions for comic panels using a VLMs pipeline without additional training.

Our contributions include the release of the code, the dense captioning test set, and the ComiCap dataset to the research community, aimed at advancing automatic comics analysis for People with Visual Impairments (PVI) and the Blind community.

Future Work. There are several avenues for future research. First, improving the accuracy of attribute extraction and text grounding by integrating more sophisticated models or combining multiple VLMs could yield better results. Second, expanding the dataset to include a wider variety of comic styles and languages could enhance the robustness and generalizability of the pipeline. Third, incorporating user feedback, especially from the PVI and Blind community, could help refine the captions to better meet their needs. Finally, exploring real-time captioning systems and their applications in educational and recreational contexts for visually impaired individuals would be a significant step forward.



In a dimly lit setting, **two men** are engaged in a tense conversation. The first man, **Atehoff**, appears startled with a frightened expression, **dressed in a blue cap** and a **dark coat** over a **red vest**. He has **graying hair** and a **weathered face**, suggesting he has seen hardships. The second man, wearing a **stylish dark suit** with a **white shirt** and a **dark tie**, has a serious demeanor, leaning slightly toward **Atehoff** while pointing, indicating authority or urgency. The background features **wooden bars** or a gate, intensifying the feeling of confinement or urgency in their dialogue.



In a **cluttered, sparsely furnished room**, **Joe Noon** confronts **two men**. The room features a **yellow backdrop with minimal decor**, emphasizing the tension of the scene. Joe, **an older man** with a **serious expression**, stands in the center wearing a **brown plaid jacket** and **holding something**, pointing at his visitors. One of the visitors, **a man with slicked-back hair** and a **surprised expression**, speaks to Joe with a hint of mockery in his voice, addressing him as "Joe". The second visitor, wearing a **dark hat** and appearing more composed, is seated and adopts a defensive posture, holding a **revolver** and suggesting he is wary of Joe's actions. The tension is palpable, highlighted by Joe's aggressive stance, suggesting a confrontation that's about to escalate.



In a vibrant comic panel, **a diverse group of people** stands excitedly, looking up towards the sky. They are all facing the same direction, **their arms raised** in a mixture of greeting and celebration. Prominent among them is **a man in a bright yellow-green outfit**, who is enthusiastically waving. Next to him, a figure in an **orange shirt** raises **his arm high**, exclaiming with joy. In the background, you can see **a couple of planes flying by**, emphasizing the festive atmosphere. The expressions of the characters are animated, capturing a moment of collective excitement as they cheer for their leader, indicated by the text characters shouting, **"HE COMES! EL GENERAL!"**, **"BRAVO!"**. The setting conveys a sense of camaraderie and anticipation, with the characters deeply engaged in the unfolding event.

Fig. 6: Qualitative results of our pipeline applied to the DCM-13k dataset for the creation of the ComiCap Dataset.

References

1. Alayrac, J.B., et al.: Flamingo: a visual language model for few-shot learning (2022), <https://arxiv.org/abs/2204.14198>
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering (2018), <https://arxiv.org/abs/1707.07998>
3. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein, J., Lavie, A., Lin, C.Y., Voss, C. (eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), <https://aclanthology.org/W05-0909>
4. Beyer, L., et al.: Paligemma: A versatile 3b vlm for transfer (2024), <https://arxiv.org/abs/2407.07726>
5. Caffagni, D., Cocchi, F., Barsellotti, L., Moratelli, N., Sarto, S., Baraldi, L., Baraldi, L., Cornia, M., Cucchiara, R.: The revolution of multimodal large language models: A survey (2024), <https://arxiv.org/abs/2402.12451>
6. Dubey, A., et al.: The llama 3 herd of models (2024), <https://arxiv.org/abs/2407.21783>

7. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation (2014), <https://arxiv.org/abs/1311.2524>
8. Iyyer, M., Manjunatha, V., Guha, A., Vyas, Y., Boyd-Graber, J., au2, H.D.I., Davis, L.: The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives (2017), <https://arxiv.org/abs/1611.05118>
9. Jiang, A.Q., et al.: Mistral 7b (2023), <https://arxiv.org/abs/2310.06825>
10. Johnson, J., Karpathy, A., Fei-Fei, L.: Densecap: Fully convolutional localization networks for dense captioning (2015), <https://arxiv.org/abs/1511.07571>
11. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos (2017), <https://arxiv.org/abs/1705.00754>
12. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Li, F.F.: Visual genome: Connecting language and vision using crowdsourced dense image annotations (2016), <https://arxiv.org/abs/1602.07332>
13. Laurençon, H., et al.: Obelics: An open web-scale filtered dataset of interleaved image-text documents (2023), <https://arxiv.org/abs/2306.16527>
14. Laurençon, H., Tronchon, L., Cord, M., Sanh, V.: What matters when building vision-language models? (2024), <https://arxiv.org/abs/2405.02246>
15. Li, Y., Aizawa, K., Matsui, Y.: Manga109Dialog A Large-scale Dialogue Dataset for Comics Speaker Detection. arXiv. <https://doi.org/10.48550/arXiv.2306.17469>, <http://arxiv.org/abs/2306.17469>
16. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2024), <https://arxiv.org/abs/2310.03744>
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023), <https://arxiv.org/abs/2304.08485>
18. OpenBMB: Minicpm-v: A gpt-4v level multimodal llm on your phone. <https://github.com/OpenBMB/MiniCPM-V> (2023)
19. Ramaprasad, R.: Comics for everyone: Generating accessible text descriptions for comic strips, <https://arxiv.org/abs/2310.00698>
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks (2016), <https://arxiv.org/abs/1506.01497>
21. Rigaud, C., Burie, J.C., Petit, S.: Toward accessible comics for blind and low vision readers (2024), <https://arxiv.org/abs/2407.08248>
22. Sachdeva, R., Shin, G., Zisserman, A.: Tails tell tales: Chapter-wide manga transcriptions with character names (2024), <https://arxiv.org/abs/2408.00298>
23. Sachdeva, R., Zisserman, A.: The manga whisperer: Automatically generating transcriptions for comics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12967–12976 (2024)
24. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Integrated recognition, localization and detection using convolutional networks (2014), <https://arxiv.org/abs/1312.6229>
25. Team, G., et al.: Gemma: Open models based on gemini research and technology (2024), <https://arxiv.org/abs/2403.08295>
26. Topal, B.B., Yuret, D., Sezgin, T.M.: Domain-adaptive self-supervised pre-training for face and body detection in drawings (2023), <https://arxiv.org/abs/2211.10641>
27. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator (2015), <https://arxiv.org/abs/1411.4555>
28. Vivoli, E., Bertini, M., Karatzas, D.: Comix: A comprehensive benchmark for multi-task comic understanding (2024), <https://arxiv.org/abs/2407.03550>

29. Vivoli, E., Campaioli, I., Nardoni, M., Biondi, N., Bertini, M., Karatzas, D.: Comics datasets framework: Mix of comics datasets for detection benchmarking (2024), <https://arxiv.org/abs/2407.03540>
30. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a unified representation for a variety of vision tasks (2023), <https://arxiv.org/abs/2311.06242>
31. Xu, R., Yao, Y., Guo, Z., Cui, J., Ni, Z., Ge, C., Chua, T.S., Liu, Z., Sun, M., Huang, G.: Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images (2024), <https://arxiv.org/abs/2403.11703>
32. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training (2023), <https://arxiv.org/abs/2303.15343>
33. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigt-4: Enhancing vision-language understanding with advanced large language models (2023), <https://arxiv.org/abs/2304.10592>