AI for Visual Art workshop

CoMiCap A VLMs PIPELINE FOR DENSE CAPTIONING OF COMICS PANELS

Vivoli Emanuele[1,2], Biondi Niccolò[2], Bertini Marco[2], Karatzas Dimosthenis[1]

[1] **Computer Vision Center**, UAB, Barcelona (ES)
[2] **MICC**, University of Florence, Firenze (IT)

paper

Github (soon)

Motivation

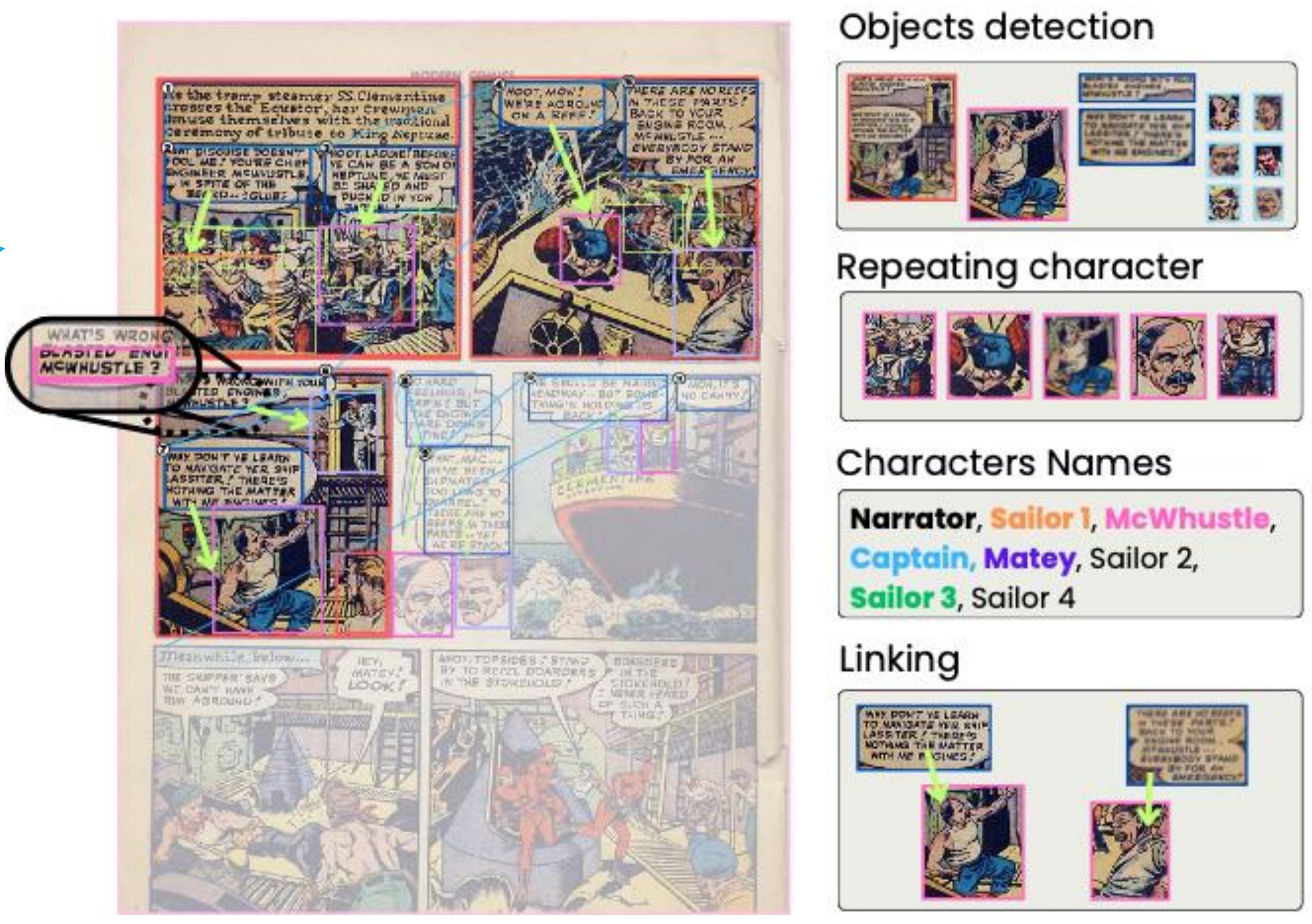
THE TASK

- Comics are **not accessible by PVI**
- Many works tackle **Dialog generation**
- ✗ **Context** about the story is **missing**

THE METRIC

- We are interested in **attributes**
- ✗ **NLP** metrics considers words-groups

Design choices


WHAT'S WRONG WITH YOU BLASTED ENGINES McWHUSTLE?
Objects detection
Repeating character
Characters Names
Narrator, Sailor 1, McWhustle, Captain, Matey, Sailor 2, Sailor 3, Sailor 4
Linking

Reading Order

Dialog

- Narrator: "As the tramp steamer [..]"
- Sailor 1: "THAT DISGUISE DOESN'T [..]"
- McWhustle: "HOOT, LADDIE! BEFORE [..]"
- McWhustle: "HOOT, MON! WE'RE AGROUND"
- Captain: "THERE ARE NO REEFS IN THESE [..]"

Scene and Description

- ? Character description
- ? Scene descripiton
- ? Objects and Attributes
- ? Actions

AR Metric

We propose a new metric for attribute retaining:

Algorithm 1 Attributes Retaining Metric Calculation

```
procedure CalculateARM($C$, $\tilde{C}^k$, $\tau$)
    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 $\tau$
    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.
    for each $i$ from 1 to $|P|$ do
        Extract predicted entity sets $\tilde{A}_i$ from predicted caption $\tilde{c}_i$ using model $x^k$
        Calculate BERT-score $BS(\tilde{A}_i, A_i)$ for each $A_i$
    end for
    for each $i$ from 1 to $|P|$ do
        Initialize cleaned set $\tilde{A}*_i \leftarrow \emptyset$
        Initialize cleaned set $J_i \leftarrow \emptyset$
        for each element $\tilde{a}$ in $\tilde{A}_i$ do
            if $BS(\tilde{a}, A_i) \geq \tau$ then
                Replace $\tilde{a}$ with the matching element $a \in A_i$
                Add $a$ to $\tilde{A}*_i$
            else
                Add $\tilde{a}$ to $\tilde{A}*_i$
            end if
        end for
        Calculate the Jaccard similarity $Jacc(\tilde{A}*_i, A_i)$
        Save it in $J_i$
    end for
    Return: $J_i$ for $i \in 1, \dots, |P|$.
end procedure
```

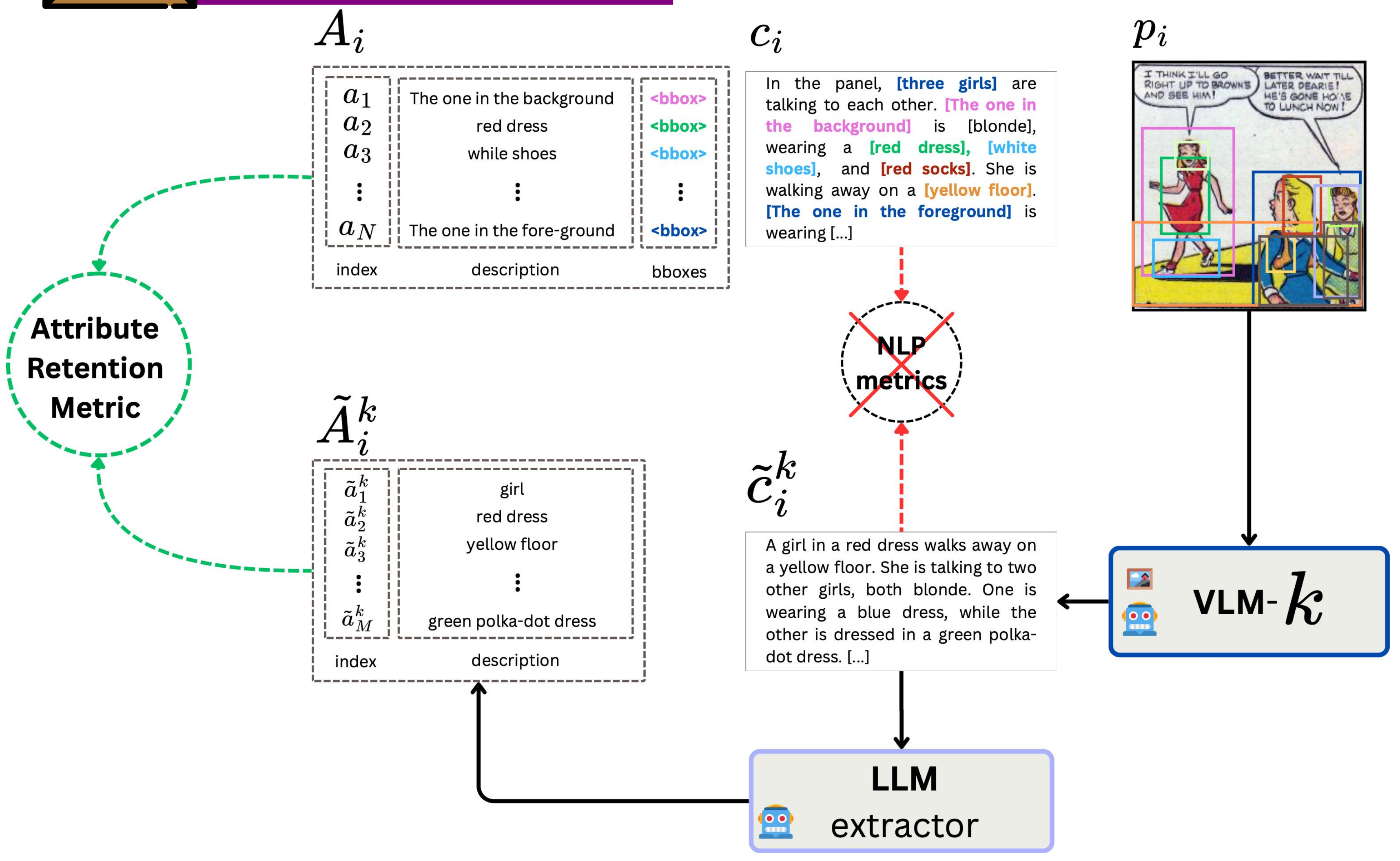
1 From a Caption ***C****

2 Extract Attributes ***A****

3 **Compare *GT* and *A**** with **BERT-score**

4 Use **IoU**

Metric



In the panel, [three girls] are talking to each other. [The one in the background] is [blonde], wearing a [red dress], [white shoes], and [red socks]. She is walking away on a [yellow floor]. [The one in the foreground] is wearing [...]

A girl in a red dress walks away on a yellow floor. She is talking to two other girls, both blonde. One is wearing a blue dress, while the other is dressed in a green polka-dot dress. [...]

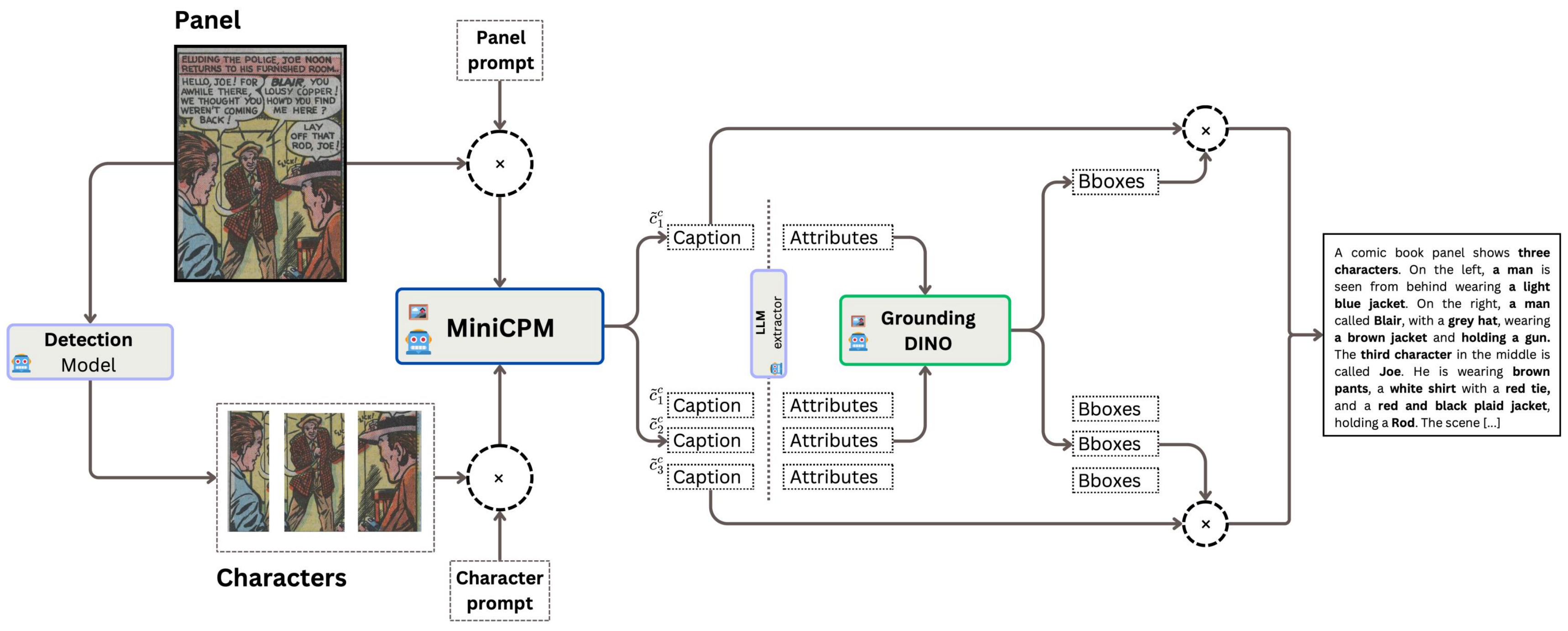
Benchmarks

We evaluate existing VLMs using our Attribute-Retention metric:

Model	ROUGE	BLEU	METEOR	ARM (ours)
PaliGemma	0.13 ± 0.02	0.01 ± <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

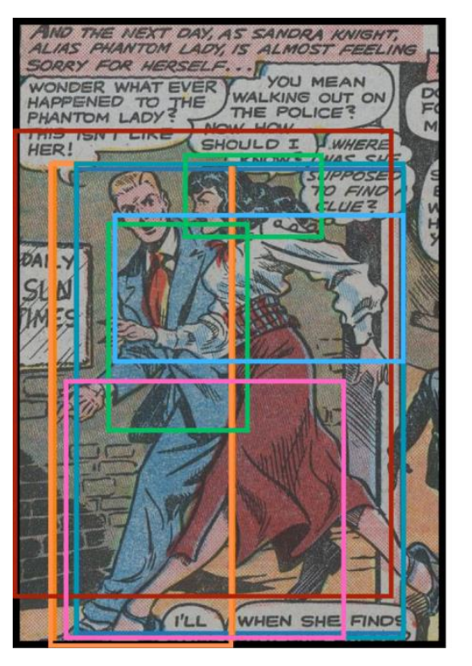
Pipeline

Full pipeline



A comic book panel shows **three characters**. On the left, **a man** is seen from behind wearing **a light blue jacket**. On the right, **a man** called **Blair**, with a **grey hat**, wearing **a brown jacket** and **holding a gun.** The **third character** in the middle is called **Joe**. He is wearing **brown pants**, a **white shirt** with a **red tie,** and a **red and black plaid jacket**, holding a **Rod**. The scene [...]

Caption & Attributes


AND THE NEXT DAY, AS SANDRA KNIGHT, ALIAS PHANTOM LADY, IS ALMOST FEELING SORRY FOR HERSELF...
WONDER WHAT EVER HAPPENED TO THE PHANTOM LADY?
YOU MEAN WALKING OUT ON THE POLICE?
I'LL WHEN SHE FINDS

$\tilde{c}^p$

~~~
```caption
 The panel depicts a scene where a young man and a woman are conversing outside a building, [..]
~~~
```csv
newspaper office
```
```

$\tilde{c}_1^c$

```
```caption
 young man, dressed in a casual blue shirt and dark pants
```
```csv
casual blue shirt
dark pants
```
```

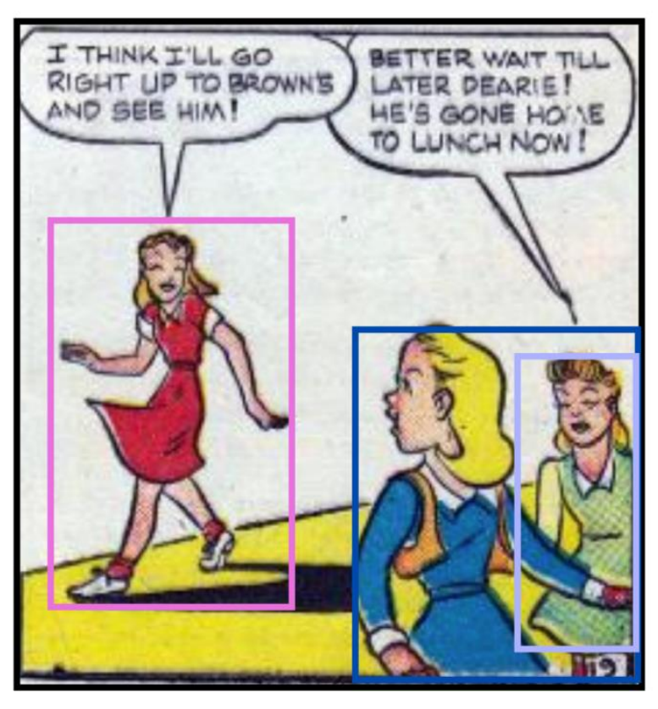
$\tilde{c}_2^c$

```
```caption
 The woman, with a dark wavy hair styled, wears a white blouse [...]
```
```csv
dark wavy hair styled
white blouse
```
```

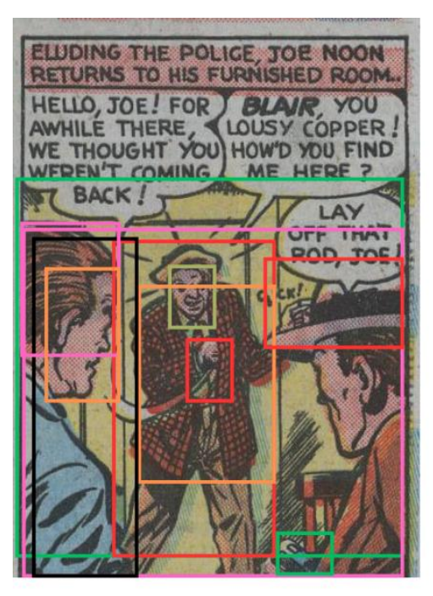
...

S&D

<attributes>	<synonyms>
white shoes	ivory footwear, ...
blonde hair	golden locks, ...
red socks	vermilion hosiery, ...


I THINK I'LL GO RIGHT UP TO BROWN'S AND SEE HIM!
BETTER WAIT TILL LATER DEARIE! HE'S GONE HOME TO LUNCH NOW!

Results


ELUDING THE POLICE, JOE NOON RETURNS TO HIS FURNISHED ROOM..
HELLO, JOE! FOR AWHILE THERE, WE THOUGHT YOU WEREN'T COMING BACK!
BLAIR, YOU LOUSY COPPER! HOW'D YOU FIND ME HERE?
LAY OFF THAT ROD, JOE!

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.

JEEPS! AIRHOFF! WHAT DO YOU WANT?
THAT DEPENDS ON WHAT BLITZ WANTS WITH THAT CONDUCTOR!

In a dimly lit setting, **two men** are engaged in a tense conversation. The first man, ~~Alehoff~~, 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 ~~Alehoff~~ while pointing, indicating authority or urgency. The background features **wooden bars** or a gate, intensifying the feeling of confinement or urgency in their dialogue.

Interested in **CoMics**?
Interested in **Vision & Language**?

Read Our Survey

survey
```