Panoptic Captioning: An Equivalence Bridge for Image and Text

NeurIPS 2025

Project Page: https://visual-ai.github.io/pancap/

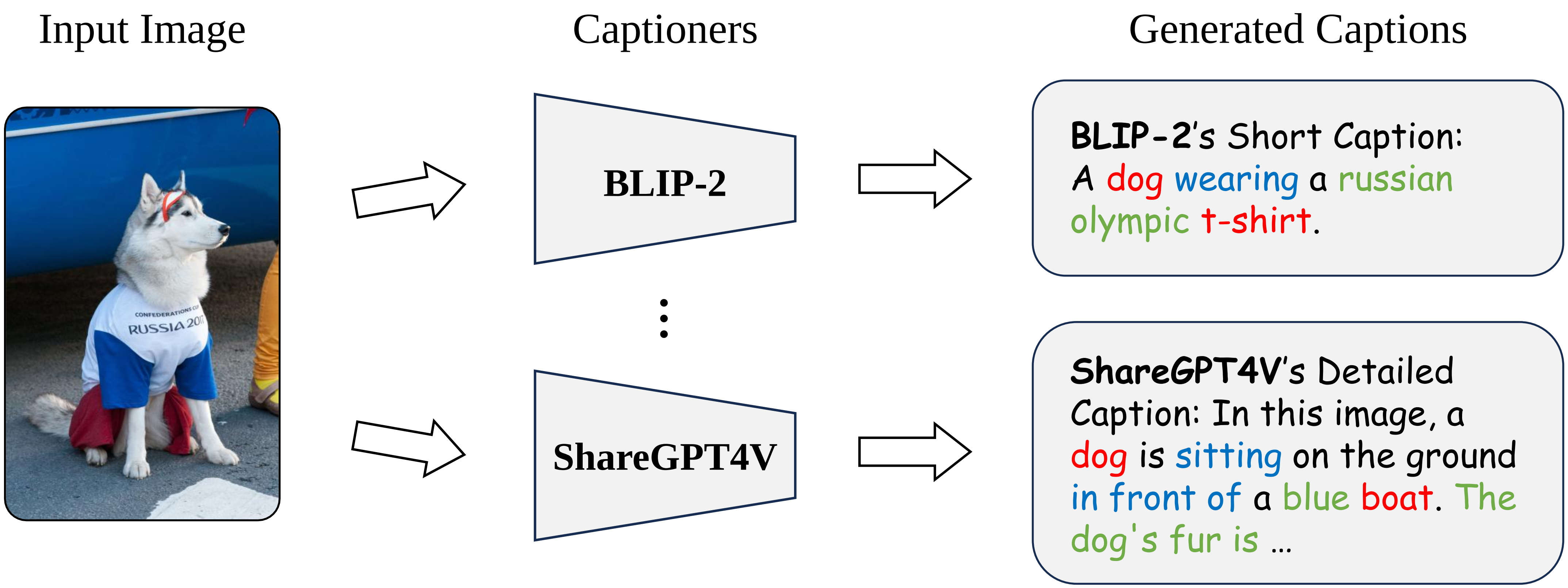
Kun-Yu Lin, Hongjun Wang, Weining Ren, Kai Han*

Visual AI Lab, The University of Hong Kong

kunyulin@hku.hk kaihanx@hku.hk

Background: Image Captioning

- Image captioning, namely representing images by textual descriptions, is a fundamental topic with broad applications.



Background: Image Captioning

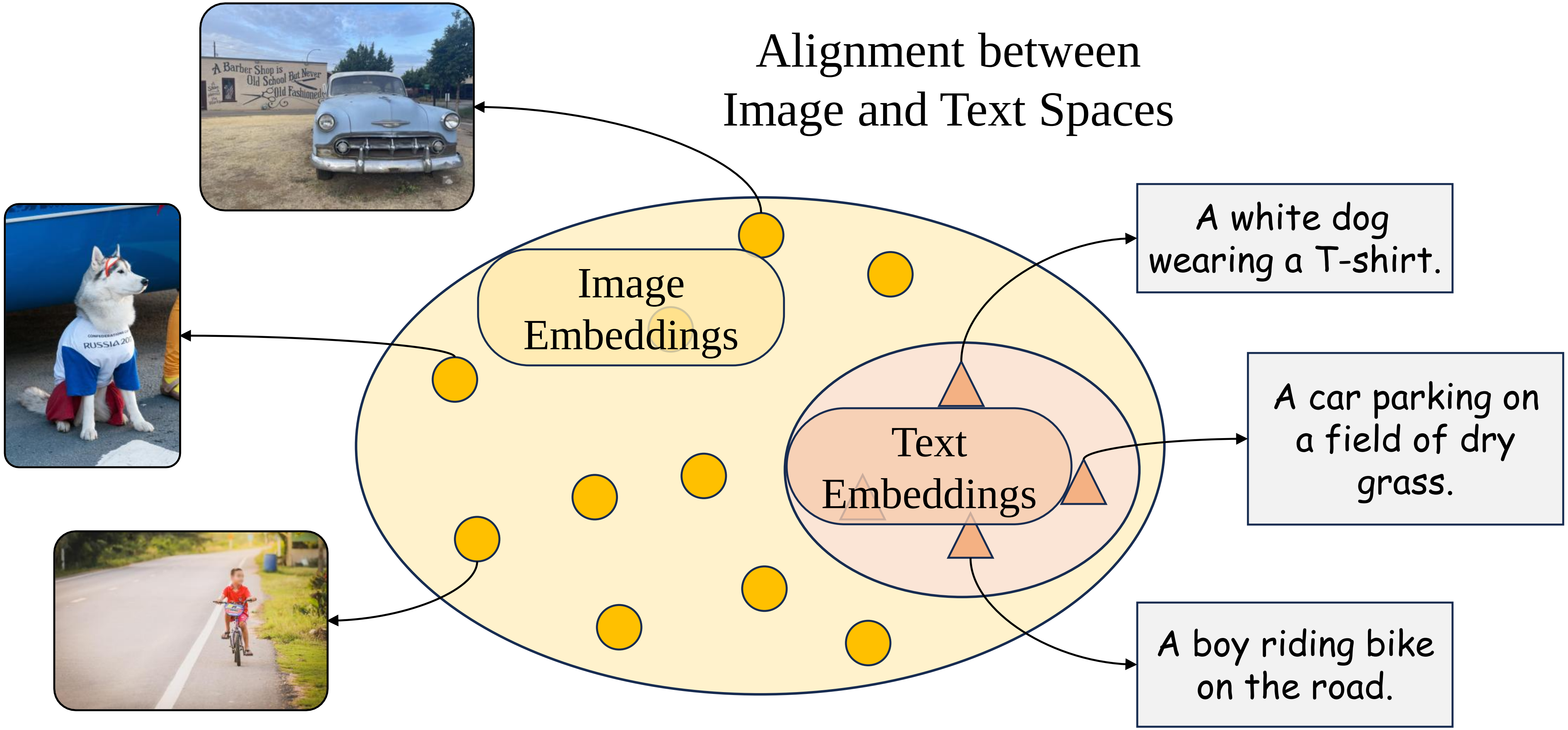
- Although existing models can produce various type of captions to describe images, their generated captions are usually too coarse, as we know "*an image is worth a thousand of words*".

Input Image

A Very Long Caption (Over 1000 words)

The image shows a husky dog sitting on the ground outdoors. It is a sunny day, with the light being even and bright, casting soft shadows, and the scene appears to be during the daytime. In the foreground, a dog, positioned at the center of the image, wears a t-shirt and a piece of fabric draped around its lower back. The dog is mostly white and gray with some black markings. It has a red and white headband around its head. Its ears are perked up, and it is looking slightly to the right. It is wearing a white t-shirt with blue sleeves. The t-shirt has writing on the front of it, which is composed of two lines of texts...

Image-Text Misalignment in Embedding Space



Concept: Image-Text Alignment in Data Space

- Our concept is to align image and text in *data* space, while existing image-text alignment models (e.g., CLIP) perform this in *embedding* space.

Input Image

Our Panoptic Caption

The image shows a husky dog sitting on the ground outdoors. It is a sunny day, with the light being even and bright, casting soft shadows, and the scene appears to be during the daytime. In the foreground, a dog, positioned at [115, 334, 1288, 2039], wears a t-shirt and a piece of fabric draped around its lower back. The dog is mostly white and gray with...

Reconstructed Image from Panoptic Caption

Concept: Image-Text Alignment in Data Space

- Our concept is to align image and text in *data* space, while existing image-text alignment models (e.g., CLIP) perform this in *embedding* space.

Input Image

Our panoptic caption serves as a bridge between the original and reconstructed images.

$\approx$

Reconstructed Image from Panoptic Caption

Panoptic Captioning: Concept and Formulation

- Our work conceives of finding the **minimum text equivalent** of an image
  - An ambitious yet challenging goal, which aims to develop a **concise** textual description that **comprehensively** captures its essential semantic elements
  - Such text representations would maximize the utility of image information for learning and downstream applications.

- This work introduces the task of **panoptic captioning**, which strives to seek the minimum text equivalent of images.
  - We formulate panoptic captioning as the task of generating a comprehensive textual description for an image, which captures all *entity instances*, their respective *locations* and *attributes*, *relationships* among entity instances, as well as *global* image state.

Panoptic Captioning: Data Example

Panoptic Caption: The image shows a husky dog sitting on the ground outdoors. It is a sunny day, with the light being even and bright, casting soft shadows, and the scene appears to be during the daytime. In the foreground, a dog, positioned at [115, 334, 1288, 2039], wears a t-shirt and a piece of fabric draped around its lower back. The dog is mostly white and gray with some black markings. It has a red and white headband around its head, and this headband is located at [850, 442, 990, 630]. Its ears are perked up, and it is looking slightly to the right. It is wearing a white t-shirt with blue sleeves. The t-shirt is within the bounding box [410, 964, 1169, 1648]. The t-shirt has writing on the front of it, which is composed of two lines of texts. The text

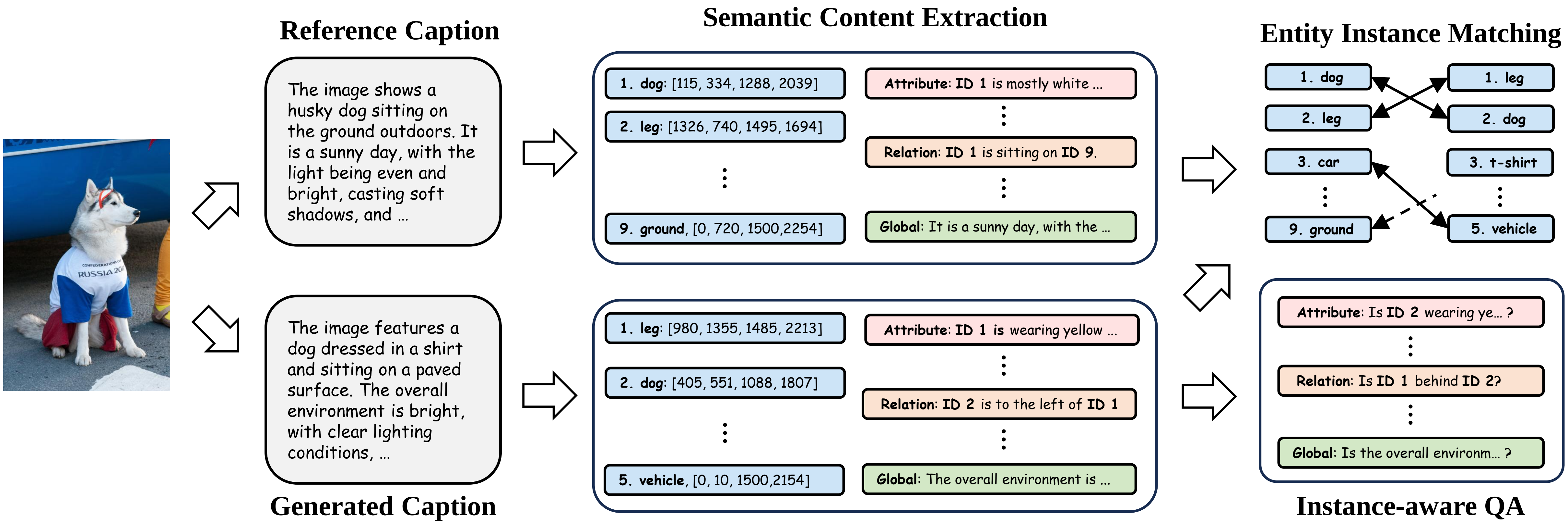
"CONFEDERATIONS CUP" appears on the t-shirt in a curved line above the text "RUSSIA 2017" in larger font. The writing's bounding box is [674, 1067, 1100, 1248]. A red fabric, positioned at [337, 1519, 1071, 1912], is possibly a makeshift pair of pants or skirt. To the right of the dog, part of a person's leg is visible, wearing orange pants, and its bounding box is [1326, 740, 1495, 1694]. The person's feet is wearing a yellow sock with a brown sandal. Only the lower leg, from just below the knee down, is visible. In the background, a part of a blue car can be seen with a bit of dark space under the vehicle. The bounding box of the car is [0, 0, 1500, 941]. The ground, positioned at [0, 720, 1500, 2254], is a gray asphalt surface. Towards the bottom-right of the image, there are white zebra markings painted on the asphalt. The markings' bounding box is [765, 2019, 1497, 2254].

Contributions

- **Task and Metric**: A novel task named panoptic captioning with a comprehensive metric, named PancapScore, for reliable evaluation.
- **Data Engine**: An effective data engine, named PancapEngine, to produce high-quality data in a detect-then-caption manner.
- **Benchmark**: A new SA-Pancap benchmark composed of high-quality auto-generated data for training and validation, and additionally provide a human-curated test set for reliable evaluation.
- **Methodology**: A simple yet effective method named PancapChain to improve panoptic captioning, which decouples the challenging panoptic captioning task into multiple subtasks

PancapScore

- The metric systematically categorizes the content into five distinct dimensions and evaluates performance on each dimension separately.



PancapEngine

- The data engine first detects diverse categories of entities in images using an elaborate entity detection suite.
  - Associate class-agnostic detection with image tagging for detecting diverse categories of entities in a given images
- We then employ state-of-the-art MLLMs to generate comprehensive panoptic captions using entity-aware prompts, ensuring the data quality by caption consistency across different MLLMs.
  - We employ Gemini-Exp-1121 to generate captions and Qwen2-VL-72B to verify data quality, due to their strong image understanding and instruction-following capabilities

The SA-Pancap Benchmark

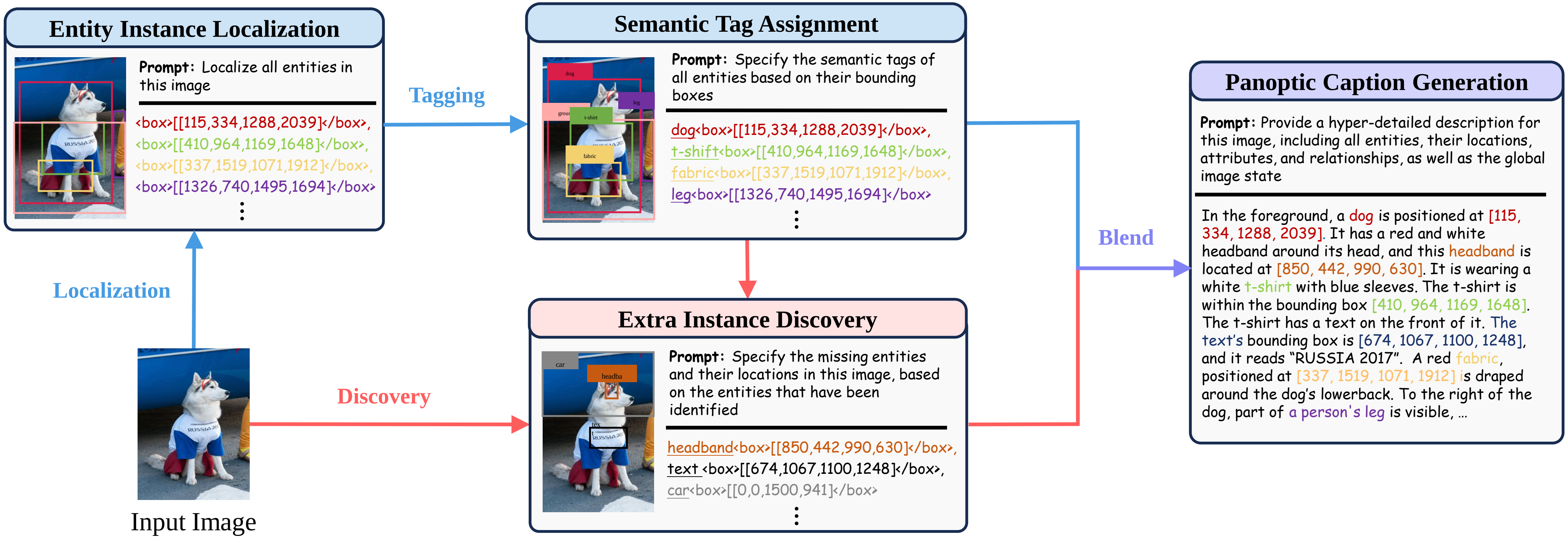
- Our SA-Pancap benchmark consists of 9,000 training and 500 validation images paired with auto-generated panoptic captions, and 130 test images paired with human-curated panoptic captions.

- Our validation and test sets consist of diverse images, paired with high-quality panoptic captions, which are selected by PancapScore.

Benchmarks	Location	Instance	Category	Sample	Token
DCI [98]	✗	-	-	7.8K	148.0
DOCCI [4]	✗	-	-	14.6K	135.7
IIW [14]	✗	-	-	9.0K	217.2
SG4V [5]	✗	-	-	1.2M	192.0
DenFu [6]	✗	-	-	1.0M	254.7
GCG [43]	✓	2.9	1329	56.9K	27.2
SA-Pancap	✓	**6.9**	**2429**	9.6K	**345.5**

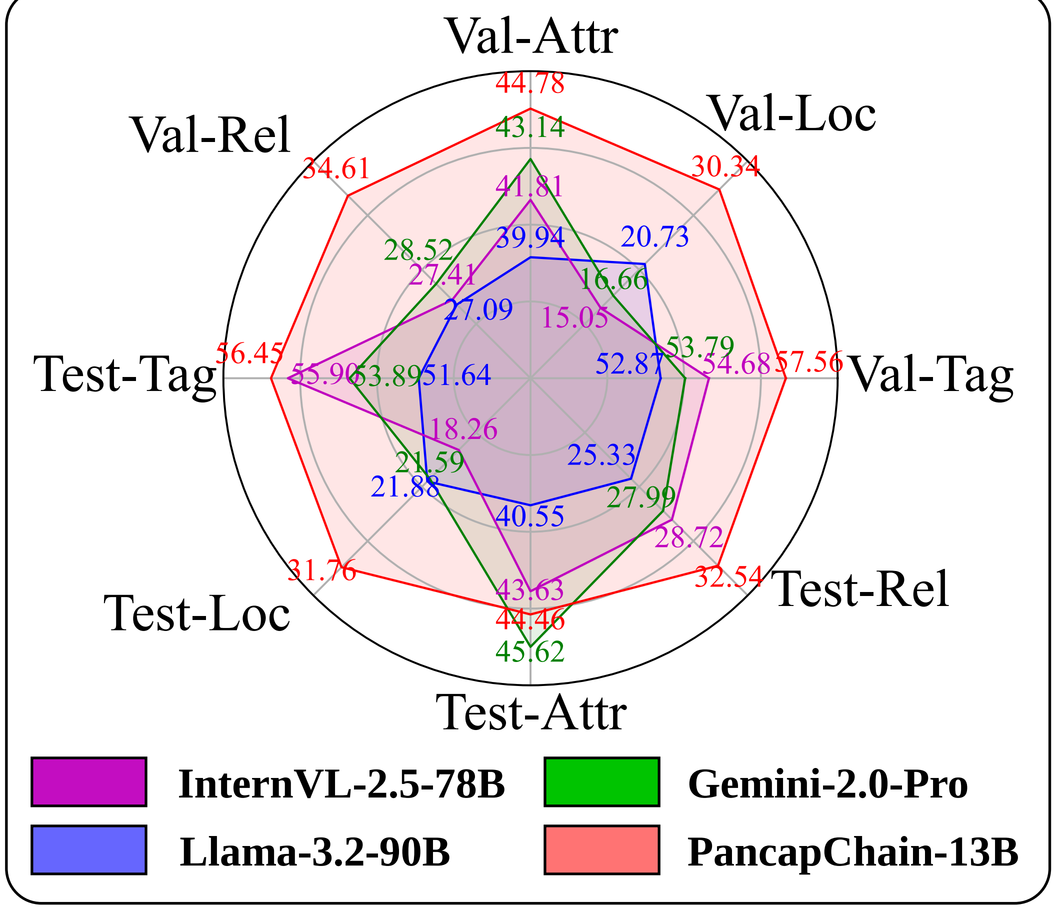
PancapChain

- Our key idea is to decouple the challenging panoptic captioning task into **multiple stages** and train the model to generate panoptic captions step by step, as an image contains rich semantic elements.



Experiment Results

- Results on SA-Pancap



- Image-Text Retrieval (DOCCI)

Comparable with SOTA Retrievers

Models	Type	R@1
CLIP [11]	Image-Text	16.9
ALIGN [16]	Image-Text	59.9
BLIP [13]	Image-Text	54.7
LongCLIP [108]	Image-Text	38.6
MATE [3]	Image-Text	**62.9**
BLIP [13]	Text-Text	47.3
ShareGPT4V [5]	Text-Text	59.6
PancapChain (Ours)	Text-Text	**61.9**

Results of “Image Reconstruction”

Thank You!