苏州大学
SOOCHOW UNIVERSITY

论文讲解

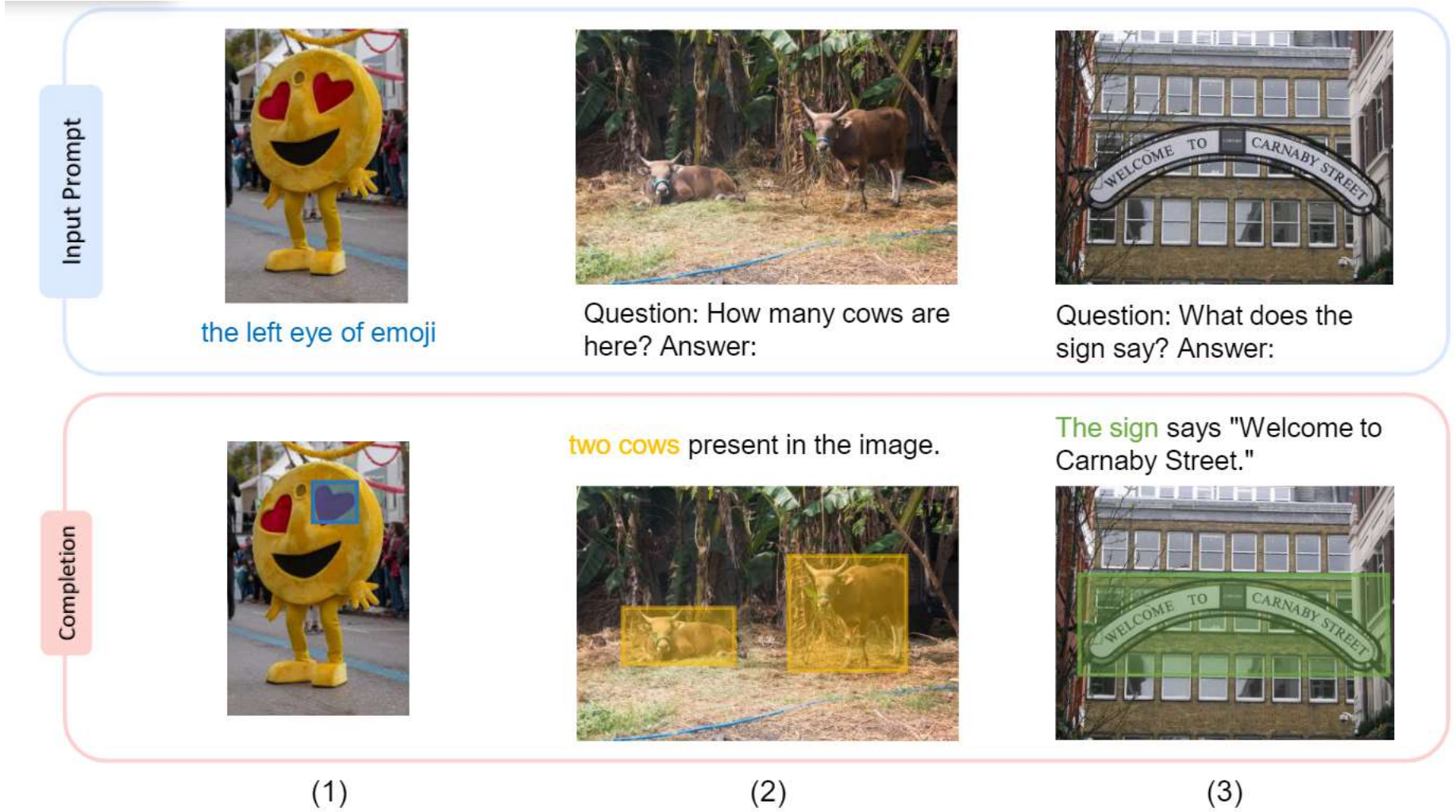
汇 报 人：陈立志

KOSMOS-2

KOSMOS-2: Grounding Multimodal Large Language Models to the World

We introduce KOSMOS-2, a Multimodal Large Language Model (MLLM), enabling new capabilities of perceiving object descriptions (e.g., bounding boxes) and grounding text to the visual world. Specifically, we represent refer expressions as links in Markdown, i.e., "[text span](bounding boxes)", where object descriptions are sequences of location tokens.

KOSMOS-2


Input Prompt
the left eye of emoji
Question: How many cows are here? Answer:
Question: What does the sign say? Answer:
Completion
two cows present in the image.
The sign says "Welcome to Carnaby Street."
WELCOME TO CARNABY STREET
(1)
(2)
(3)

KOSMOS-2

Input Prompt

Question: Why is this animal unusual? Answer:

Question: What is it? Answer:

Question: What is the biggest difference betwen bottle-1 and bottle-2? Answer:

Completion

This animal is unique because it is a turtle. Turtles are known for their slow and steady pace, whereas rabbits are known to be agile and quick. It is not common for a turtle to be faster than a rabbit, especially in terms of speed.

a buoy attached to the boat.

The biggest difference between bottle-1 and bottle-2 is the label.

(4) (5) (6)

Input Prompt

Describe this image in detail:

Completion

A snowman sits next to a campfire in the snow. He is wearing a hat, scarf, and mittens. There are several pots nearby, likely containing a hot meal for the snowman. The scene is serene and peaceful, with the snowman looking at the fire with a warm smile.

(7)

Construction of Web-Scale Grounded Image-Text Pairs

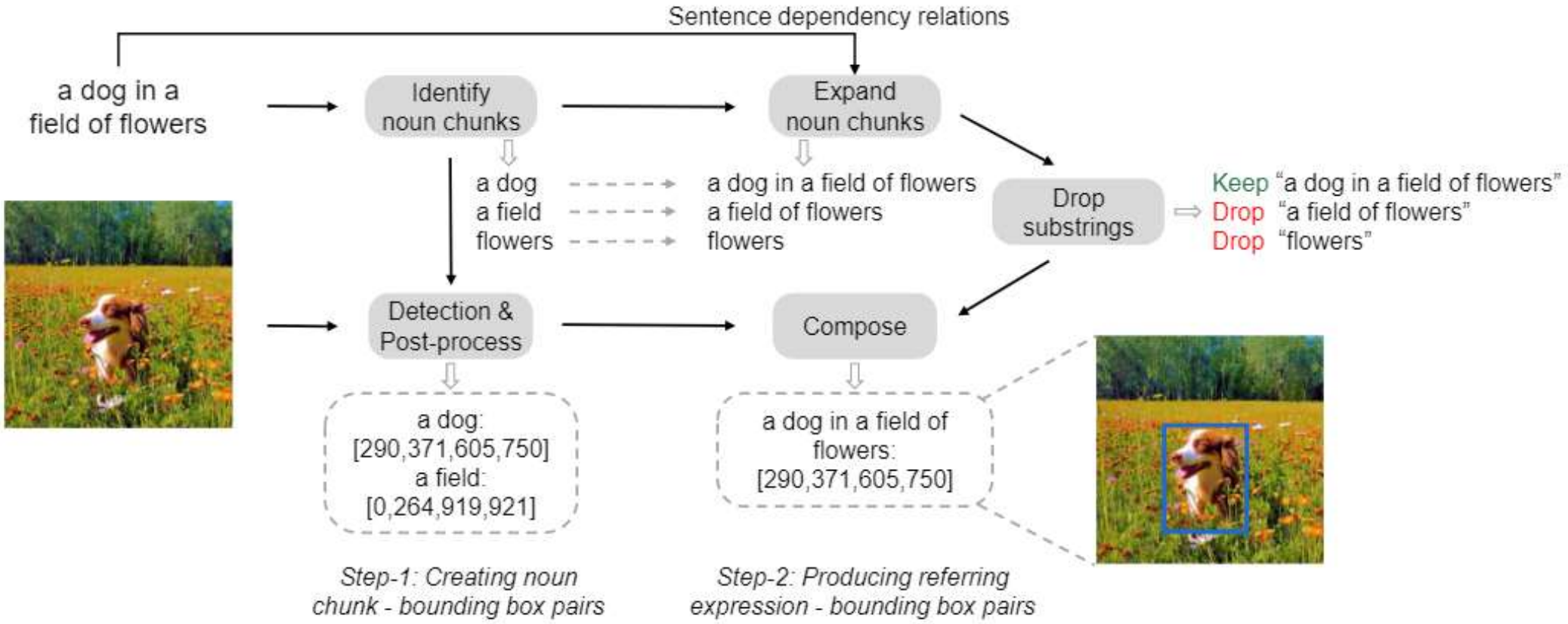


Figure 3: The pipeline of constructing web-scale grounded image-text pairs.

Input Representations

In total, P × P location tokens are introduced, and these tokens are added to word vocabulary to enable unified modeling with texts.

<s> <image> Image Embedding </image> <grounding> It <box><loc$_{44}$><loc$_{863}$></box> seats next to a campfire <box><loc$_{4}$><loc$_{1007}$></box> </s>

- "What is it <box><loc$_1$><loc$_2$></box>? It is {*expression*}."
- "What is this <box><loc$_1$><loc$_2$></box>? This is {*expression*}."
- "Describe this object <box><loc$_1$><loc$_2$></box>. This object is {*expression*}."
- " It <box><loc$_1$><loc$_2$></box> is {*expression*}."
- " This <box><loc$_1$><loc$_2$></box> is {*expression*}."
- " The object <box><loc$_1$><loc$_2$></box> is {*expression*}."

Table 9: Instruction templates used for expression generation.

COSA

COSA: Concatenated Sample Pretrained Vision-Language Foundation Model

- Due to the limited scale and quality of video-text training corpus, most vision language foundation models employ image-text datasets for pretraining and primarily focus on modeling visually semantic representations while disregarding temporal semantic representations and correlations.
- To address this issue, we propose COSA, a Concatenated Sample pretrained vision-language foundation model. COSA jointly models visual contents and event-level temporal cues using only image-text corpora.
- We achieve this by sequentially concatenating multiple image-text pairs as inputs for pretraining.

COSA

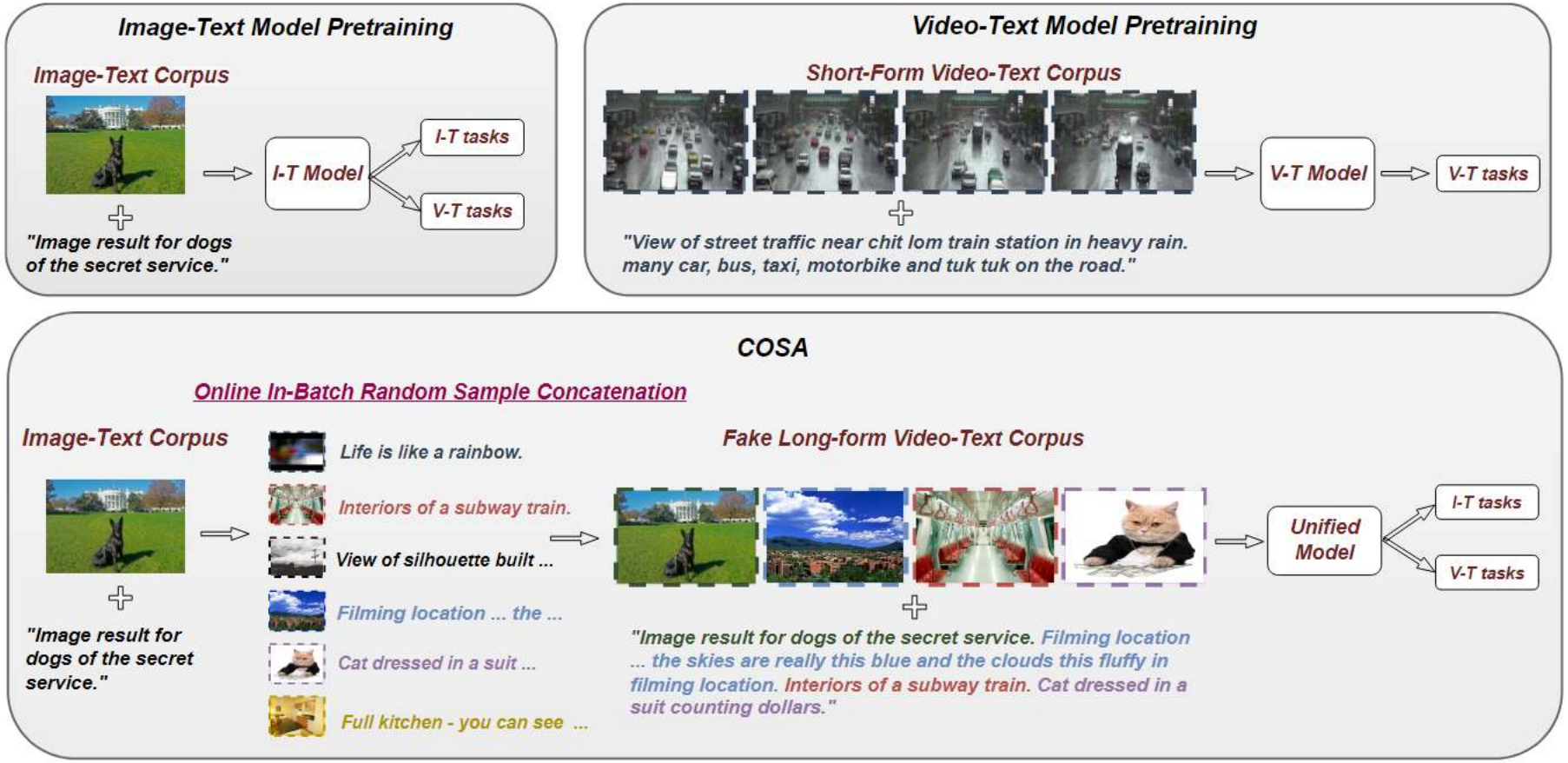


Figure 1: Visualizations of the traditional image-text and video-text model pretraining pipeline and the proposed unified COSA, which transforms the image-text corpus into a synthetic long-form video-text corpus online through random sample concatenation.

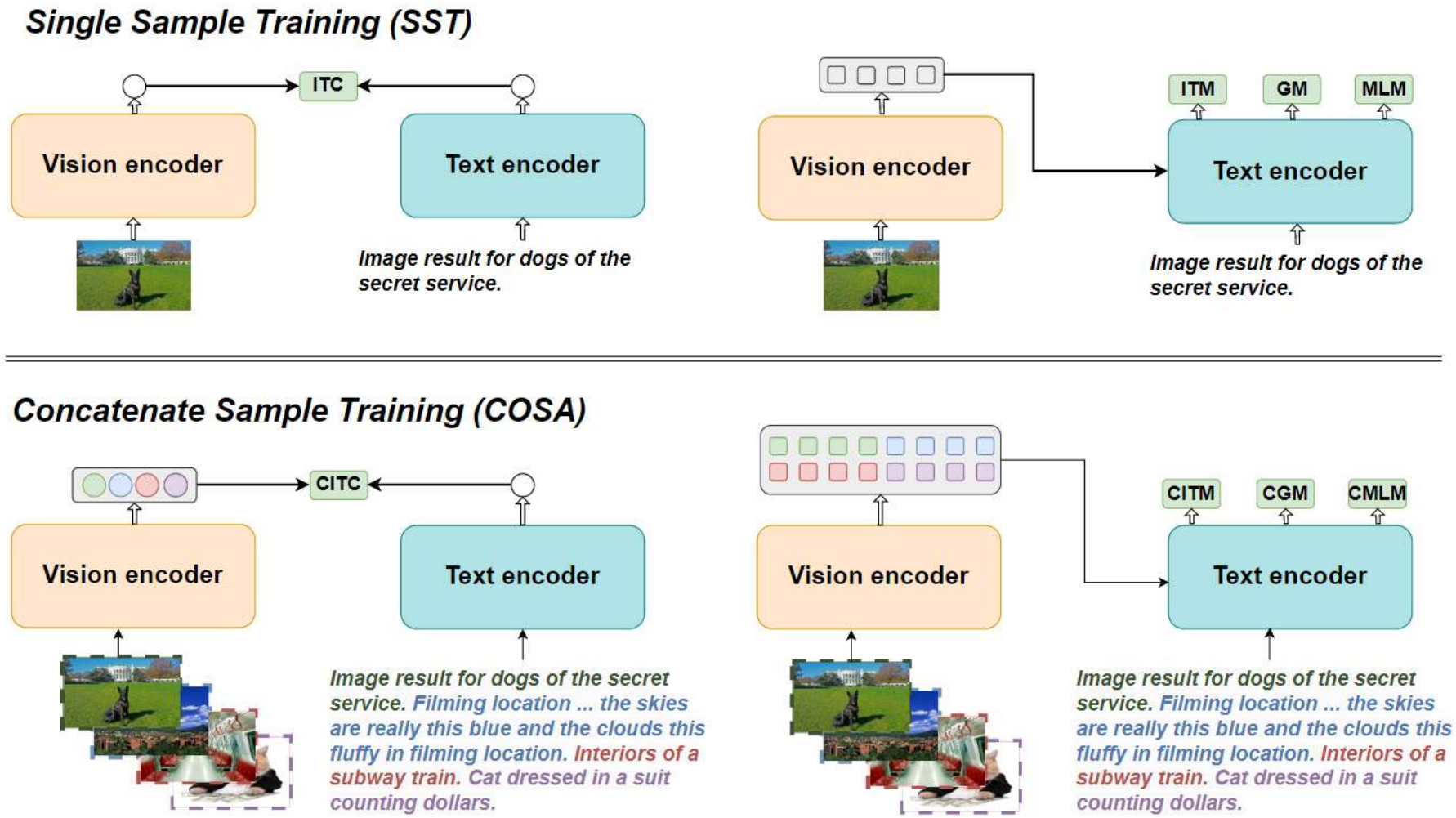


Figure 2: Visualizations of the training framework for COSA (bottom). In contrast to the conventional single sample training framework (SST), COSA takes the on-the-fly transformed pseudo long-form video-paragraph corpus as input. Circles and squares in the figure represent global and patch features, respectively.

蘇州大學
SOOCHOW UNIVERSITY

THANKS

汇 报 人：陈立志