DeCLIP: Hard Negative Mining through Image Decomposition and Inpainting

Zhanhao Liu, Qiulin Fan, Huanchen Jia, Lingyu Meng

Department of Mathematics & Department of Computer Science, University of Michigan

Motivation

Despite CLIP's success in vision-language learning, it struggles with fine-grained understanding due to its reliance on global image-text alignment, limiting performance in tasks requiring subtle visual distinctions.

How can we push CLIP to better learn localized visual details?
Our project explores *image-based hard negatives* as a way to enhance CLIP's attention to fine-grained features and improve its robustness in real-world multimodal tasks.

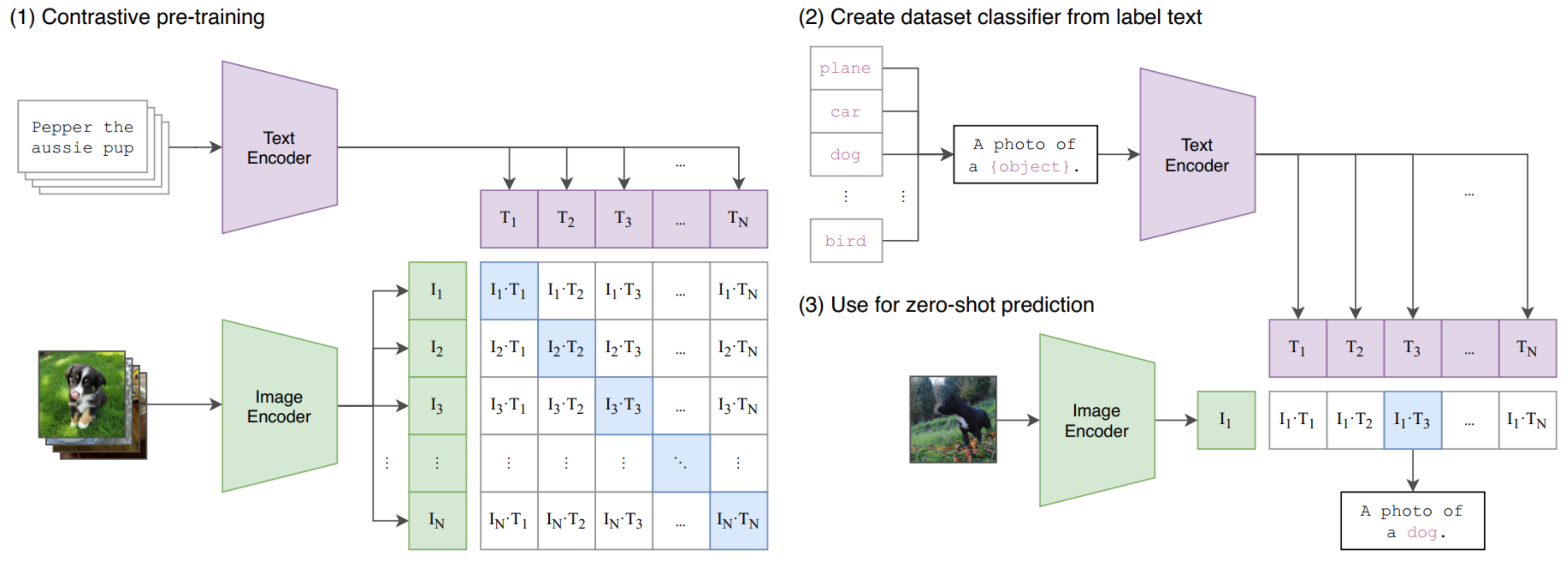
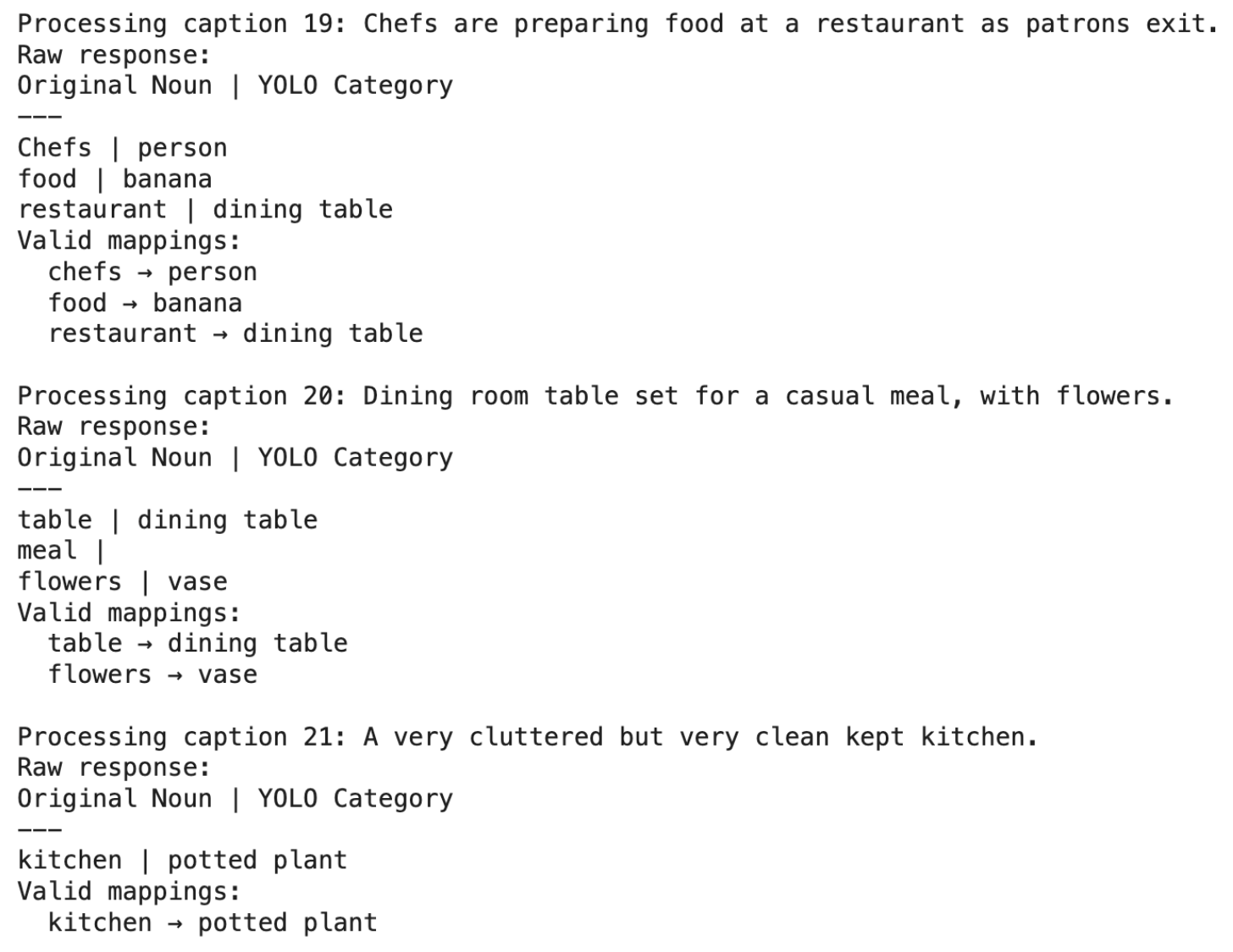

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 … TN
I1 I2 I3 … IN
I1·T1 I1·T2 I1·T3 … I1·TN
I2·T1 I2·T2 I2·T3 … I2·TN
I3·T1 I3·T2 I3·T3 … I3·TN
IN·T1 IN·T2 IN·T3 … IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
I1·T1 I1·T2 I1·T3 … I1·TN
A photo of a dog.

Figure 1. Original CLIP Model

Dataset and Preprocessing

- **Dataset:** The MS COCO 2014 dataset containing 164,000 images (82,783 of which for training), and each image is paired with 5 captions. We use the first one for keyword extration and further rewriting.
- **Preprocessing for caption (Stage 1):** To guide DeClip's visual–textual grounding, on the caption end, we preprocess each caption in two stages:
  Keyphrase Extraction: We parse each caption using a part-of-speech tagger and select the most salient nouns that denote primary objects or spatial relations, ensuring the model focuses on core scene elements.
  Category Mapping: Each extracted noun is translated into its corresponding YOLO detection category from the 80 classes defined in YOLOv8, allowing DeClip to leverage pretrained object-detection priors and attend more to the difference of corresponding regions during training.
- **Preprocessing for image and caption (Stage 2):** We then use YOLOv8 to segment the image, generating masks. For keywords that generate empty masks, it means the keyword-category map is not correct, so we remove the keyword from the map.

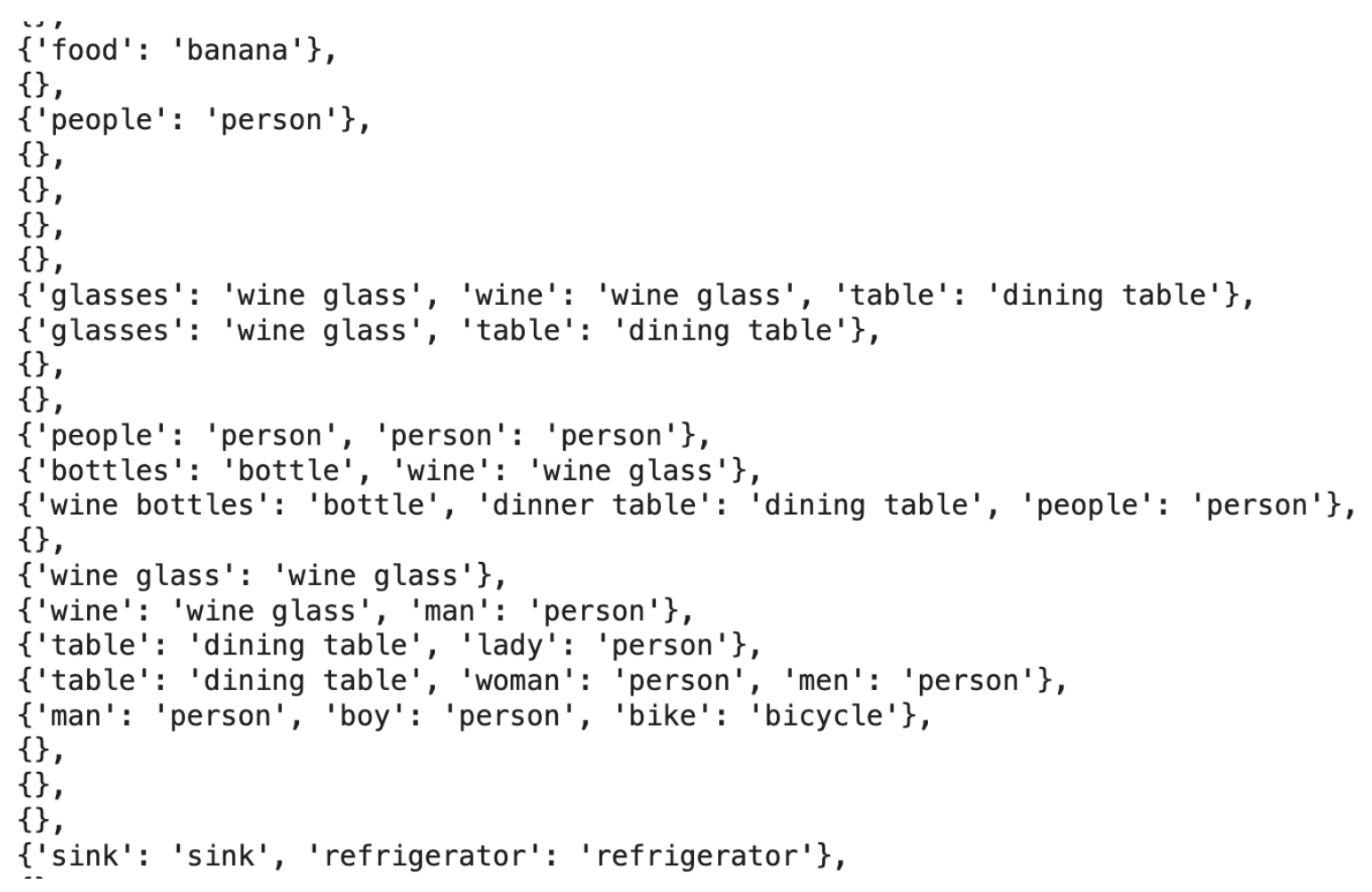


```
Processing caption 19: Chefs are preparing food at a restaurant as patrons exit.
Raw response:
Original Noun | YOLO Category
---
Chefs | person
food | banana
restaurant | dining table
Valid mappings:
  chefs → person
  food → banana
  restaurant → dining table

Processing caption 20: Dining room table set for a casual meal, with flowers.
Raw response:
Original Noun | YOLO Category
---
table | dining table
meal |
flowers | vase
Valid mappings:
  table → dining table
  flowers → vase

Processing caption 21: A very cluttered but very clean kept kitchen.
Raw response:
Original Noun | YOLO Category
---
kitchen | potted plant
Valid mappings:
  kitchen → potted plant
```

Figure 2. LLM keyword extration and mapping



```
{},
{'food': 'banana'},
{},
{'people': 'person'},
{},
{},
{},
{},
{'glasses': 'wine glass', 'wine': 'wine glass', 'table': 'dining table'},
{'glasses': 'wine glass', 'table': 'dining table'},
{},
{},
{'people': 'person', 'person': 'person'},
{'bottles': 'bottle', 'wine': 'wine glass'},
{'wine bottles': 'bottle', 'dinner table': 'dining table', 'people': 'person'},
{},
{'wine glass': 'wine glass'},
{'wine': 'wine glass', 'man': 'person'},
{'table': 'dining table', 'lady': 'person'},
{'table': 'dining table', 'woman': 'person', 'men': 'person'},
{'man': 'person', 'boy': 'person', 'bike': 'bicycle'},
{},
{},
{},
{'sink': 'sink', 'refrigerator': 'refrigerator'},
```

Figure 3. Updated keyword-category map after segmentation

Methodology

We propose **DeCLIP**, a framework that improves fine-grained visual understanding by generating **hard negatives** on the *image side*.
Given an image-text pair:

- Remove key objects from the image to create a **dual image**, keeping the caption unchanged → a semantically mismatched hard negative.
- Remove corresponding keywords from the caption to create a **dual text** aligned with the altered image → a new positive pair.

Surely, dual images and original captions form false examples and almost surely, dual images and dual captions form negative examples.

Dual Image and Dual Text Generation

1. **Diffusion Inpainting:** Remove masked objects to generate **dual images**.
2. **Caption Rewriting:** Use LLM to remove matched keywords, producing **dual texts**.

Theoretical Motivation. Based on the *manifold hypothesis*, we assume:

- Dual images/texts lie near originals in embedding space.
- This helps CLIP learn subtle mismatches and improves multimodal robustness.

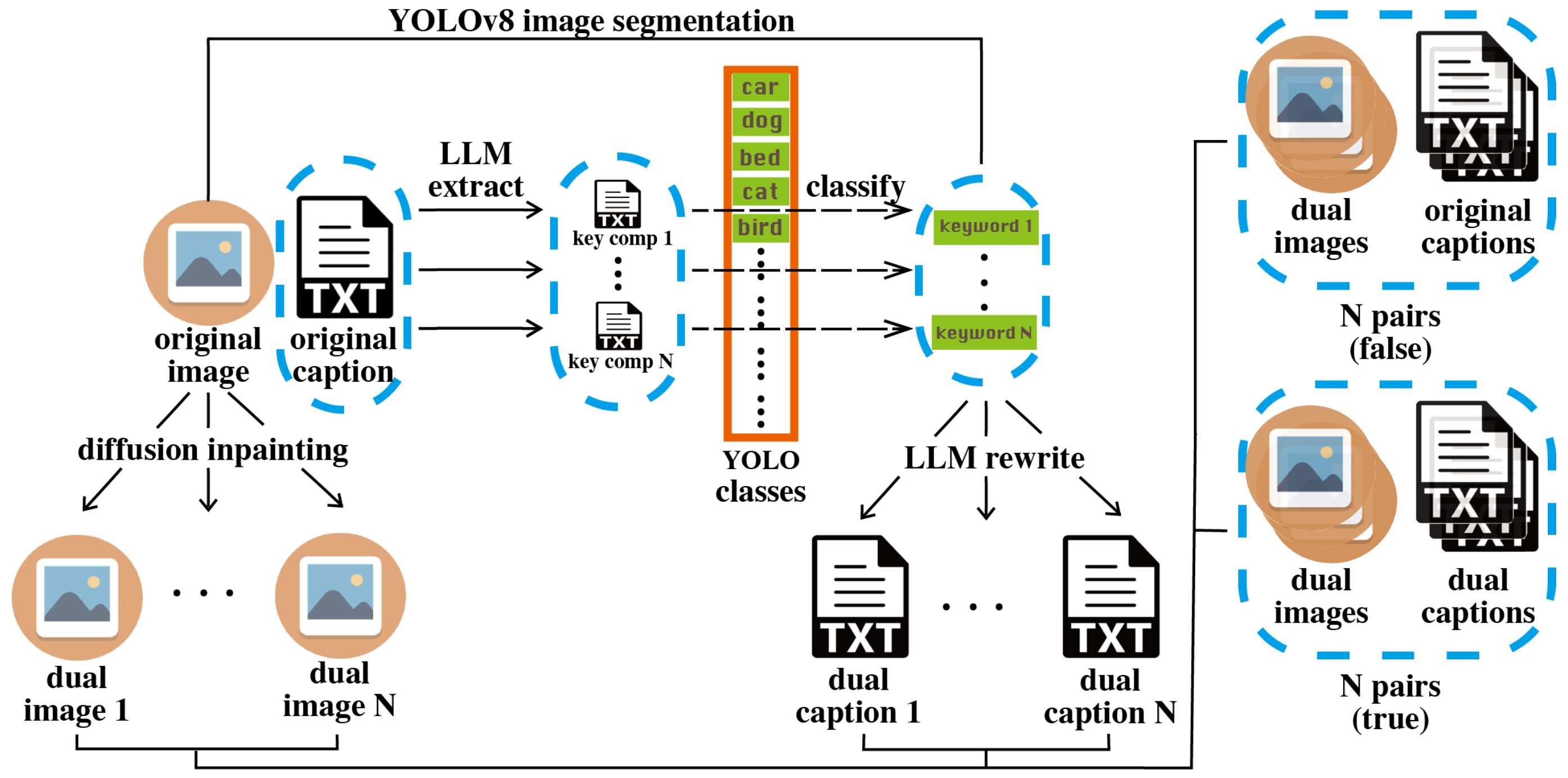


Figure 4. Conceptual diagram of DeCLIP pipeline

Training and Fine-tuning

In this section, we evaluate our model using a **margin-based loss function**, which is an enhancement of the original CLIP loss. The loss for each example is defined as:

$$\ell = \frac{1}{N}\sum_{j=1}^{N} \max\left(0,\ \text{MARGIN} + s_j^- - s^+\right)$$

This formulation incorporates a margin term to enforce a minimum difference between positive and negative example scores. The use of the ReLU activation function ensures that only violations of this margin contribute to the loss. By averaging over all negative samples, the loss function emphasizes positive examples, encouraging the model to learn more discriminative features.

Result

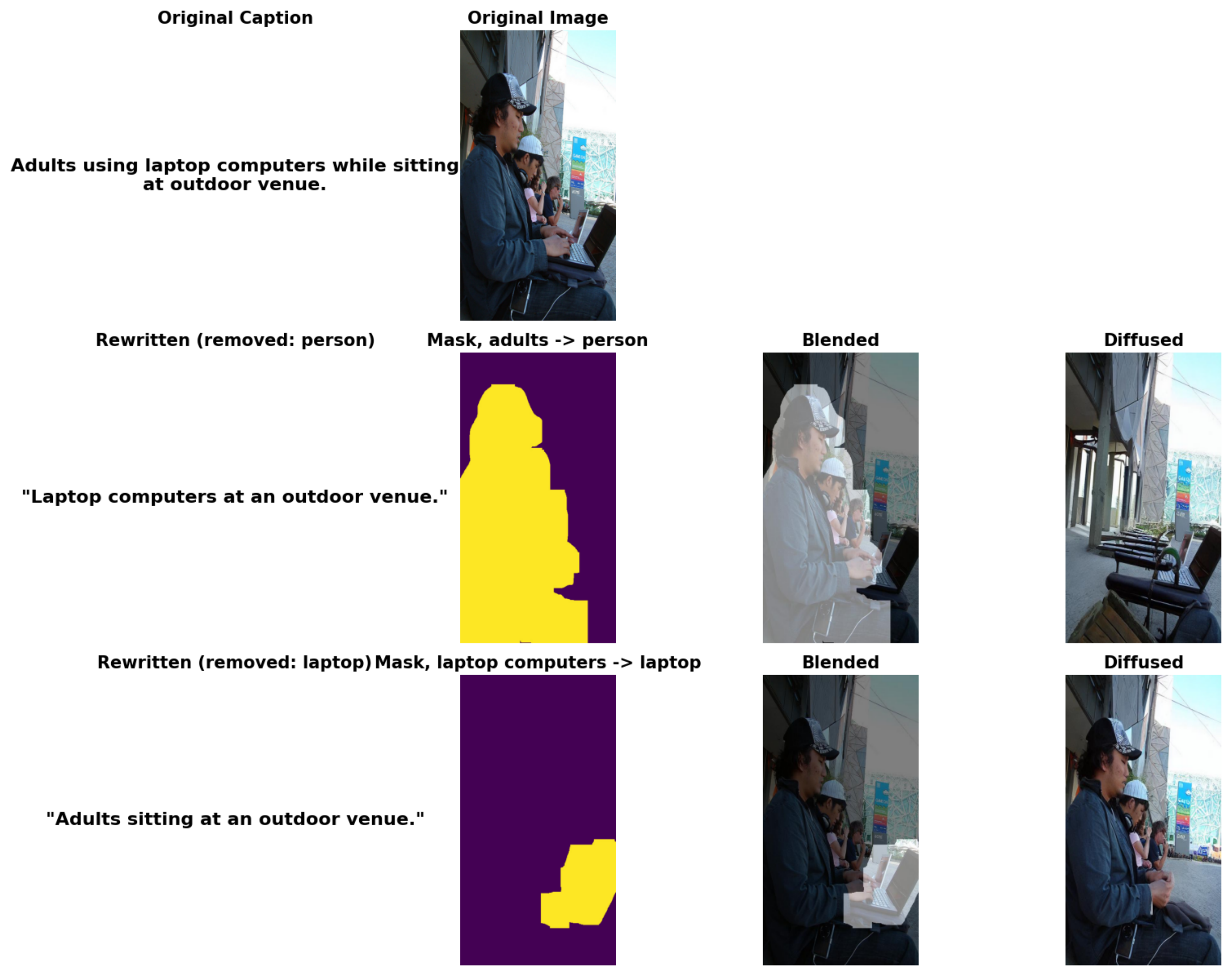

Original Caption
Original Image
Adults using laptop computers while sitting at outdoor venue.
Rewritten (removed: person)
Mask, adults -> person
Blended
Diffused
"Laptop computers at an outdoor venue."
Rewritten (removed: laptop)
Mask, laptop computers -> laptop
Blended
Diffused
"Adults sitting at an outdoor venue."

Figure 5. A random sample of dual (image, caption) pairs generated by Declip.

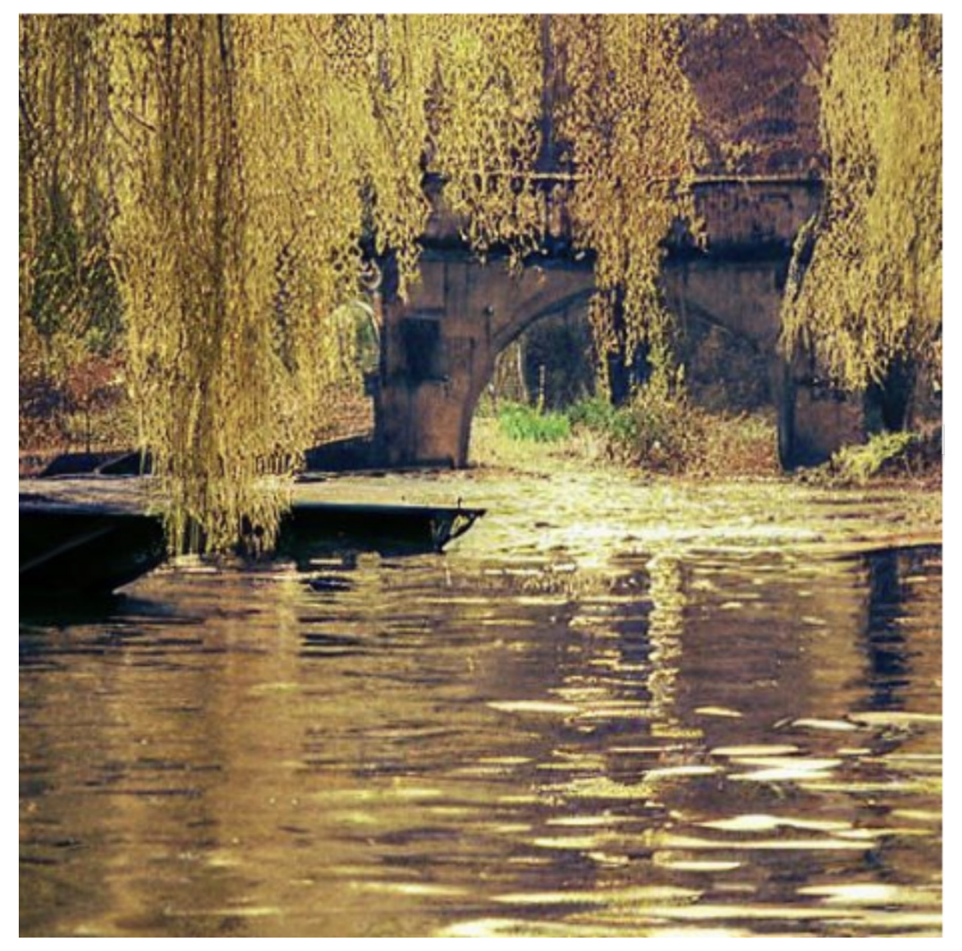


Image 1: Removing People

Image 2: Original

Image 1 similarity (Original CLIP): 0.3248, Image 1 similarity (DeCLIP): **0.4197**
Image 2 similarity (Original CLIP): 0.2788, Image 2 similarity (DeCLIP): **0.1250**

Ongoing and Future Work

Now we have automated the process and finished generation of dual imgaes and dual captions of 5,000 training examples together with correspondng fine-tune training of DeClip. Current trend shows that. The resulting dual images and texts enrich training with both false and new positive pairs, boosting model robustness.

Ongoing

- **Scalability And Efficiency:** Keep on generating dual images and texts for the whole COCO dataset, and optimize the pipeline for large-scale datasets (e.g., 3M images), balancing computational cost and inpainting quality.
- **Benchmark:** Soon after we finished the process of the whole COCO dataset, We will evaluate DeCLIP's zero-shot capability on small-scale datasets such as CIFAR-100 and STL-10, as well as on ImageNet; and compositional reasoning ability using the ARO benchmark.(Till today, we have already seen a bit improvement.)