Relational Visual Similarity

Supplementary Material

7. Implementation Details

This section presents implementation details as well as snapshots of the training data and model predictions. For specific details about the hyperparameters, etc., please visit our GitHub repository and the Hugging Face Datasets page.

Interesting images filtering prompt

You are an expert in visual creativity and interestingness. Your task is to determine if the given image is visually interesting or not.
If the image is interesting, answer "Yes".
If the image is not interesting, answer "No".
Remember, you are only allowed to answer "Yes" or "No", no other words or phrases.

Interesting Image Filtering. We trained an image filtering model on 1.3k positive images and 11k negative images. The model used was Qwen2.5-VL-7B-Instruct [41], trained with LoRA. Positive images were labeled as "Yes" (the model should answer "Yes"), and negative images were labeled as "No" (the model should answer "No") accordingly. Examples of images classified as positive and negative are shown in Fig. 12. The keep rate is around 0.7% (i.e., out of every 1k images, the model marks about 7 as "interesting").

Write anonymous caption for each image prompt

You are given a single image.
Carefully analyze it to understand its underlying logic, layout, structure, or creative concept. Then generate a single, reusable anonymous caption that could describe any image following the same concept.
The caption must:
- Fully capture the general logic or analogy of the image.
- Include placeholders (e.g., {Object}, {Word}, {Character}, {Meaning}, {Color}, etc.) wherever variations can occur.
- Be concise and standalone.

Important: Only output the anonymous caption. Do not provide any explanations or additional text.

Anonymous captioning model. The full prompt for obtaining the anonymous captions for each image group, and the prompt used to train the anonymous captioning model, are provided below. We also present an example of a predicted caption for each image in Fig. 13.

*"Artwork depicting **{Animal}** with the features and traits of a **{Historical Figure}**."*

*"Creative clothing design featuring an exaggerated representation of a **{Object}** with vibrant colors and detailed patterns."*

*"Creative arrangement of **{Animals}** made with **{Puzzle Pieces}** showcasing real-life counterparts."*

*"A monochrome image with one **{Object}** highlighted in bright **{Color}**."*

*"Intriguing images of **{Animal 1}** and **{Animal 2}** forming a unique bond."*

Figure 13. Example of predicted anonymous caption

Anonymous captions for image group

You are given two or more images that share a common logic, layout, structure, or creative concept (e.g., alphabet worksheets, step-by-step drawings, animals made from peeled fruits, etc.).
Your task is to carefully analyze all the images, identify the shared logic or analogy among them, and create one anonymous caption that describes all the images.
The anonymous caption must:
- Be a single, reusable image caption that fully describes the general logic of all the images.
- Must include placeholders (e.g., {Object}, {Word}, {Character}, {Meaning}, {Color}, etc.) wherever variations occur.

For example: "Image of using {Fruit} to create a {Animal}"; "Growth process of {Subject} described in 4 main stages: {Stage 1}, {Stage 2}, {Stage 3}, {Stage 4}"
Only provide the anonymous caption; Do not include any other explanation or content.

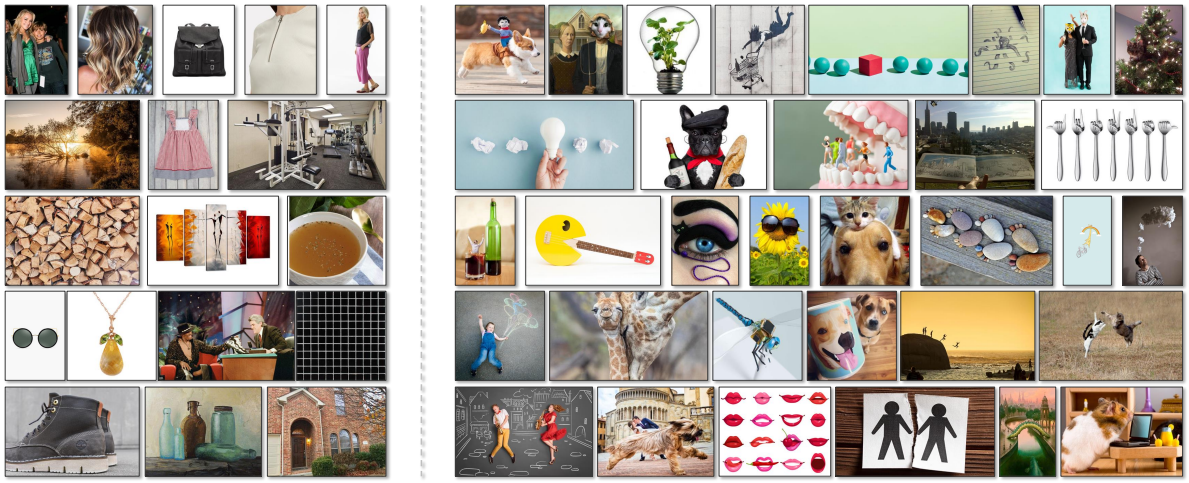

Figure 12. Examples of interesting and uninteresting images filtered by the finetuned Image Filtering model.

Automated Judgment. We present the full prompt used for automated judgment of a query image and a retrieved image below.

Automated Judgment for Image Retrieval

You are given two images.
Your task is to determine whether these two images share a similar underlying logic—that is, whether they form an analogical pair.
Do NOT base your judgment on visual similarity (e.g., color, shape, composition) or semantic similarity (such as both showing the same object or class). Images that are visually or semantically similar but do NOT share the same underlying logic should receive a very low score.
Focus ONLY on whether the two images convey the same conceptual or relational logic. For example, if one image shows a peach's internal structures, and the other shows a Earth's internal structures, they share the same logic and should receive a very high score.
Output only the number.
- 10 = very strong analogical/relational similarity (same underlying logic)
- 0 = no logical/relational similarity

Please directly output the score.

8. Additional Results

Additional image retrieval results can be found in Fig. 14-15

9. Data Collection and Annotation

The key details of data collection and annotation process are listed as below.

About annotators. All annotation instructions are written in English. All annotators are proficient in English and familiar with Computer Vision (PhD students/holders).

Annotating Interesting Images. Three annotators were shown 10 different groups of "interesting" images (e.g., 3 of them are shown in Fig. 1, Group A) and similarly 5 different groups of "not interesting" images (e.g., 3 of them are shown in Fig. Fig. 1, Group B). We randomly sampled 15k images from LAION-2B [18] and simply instructed the annotators to click to select the "interesting" images (Line 232). The agreement between these 3 annotators was 92%.

Annotating Image Groups. 532 image groups have been collected. As above, we showed 10 examples of "interesting groups" (e.g., Fig. Fig. 1, Group A) to nine annotators and asked them to manually find and propose additional groups. In total, ∼400 groups were proposed. All proposed groups were further verified by three annotators, and we retained only those for which all annotators agreed that there was a clear, non-duplicate pattern (85% of the groups was kept).

Data Attributions

All images used in this paper are from the publicly available LAION-2B dataset [18]. The authors do not own any of the images and acknowledge the dataset creators and/or the original copyright holders of each image. All images are used for research purposes only.

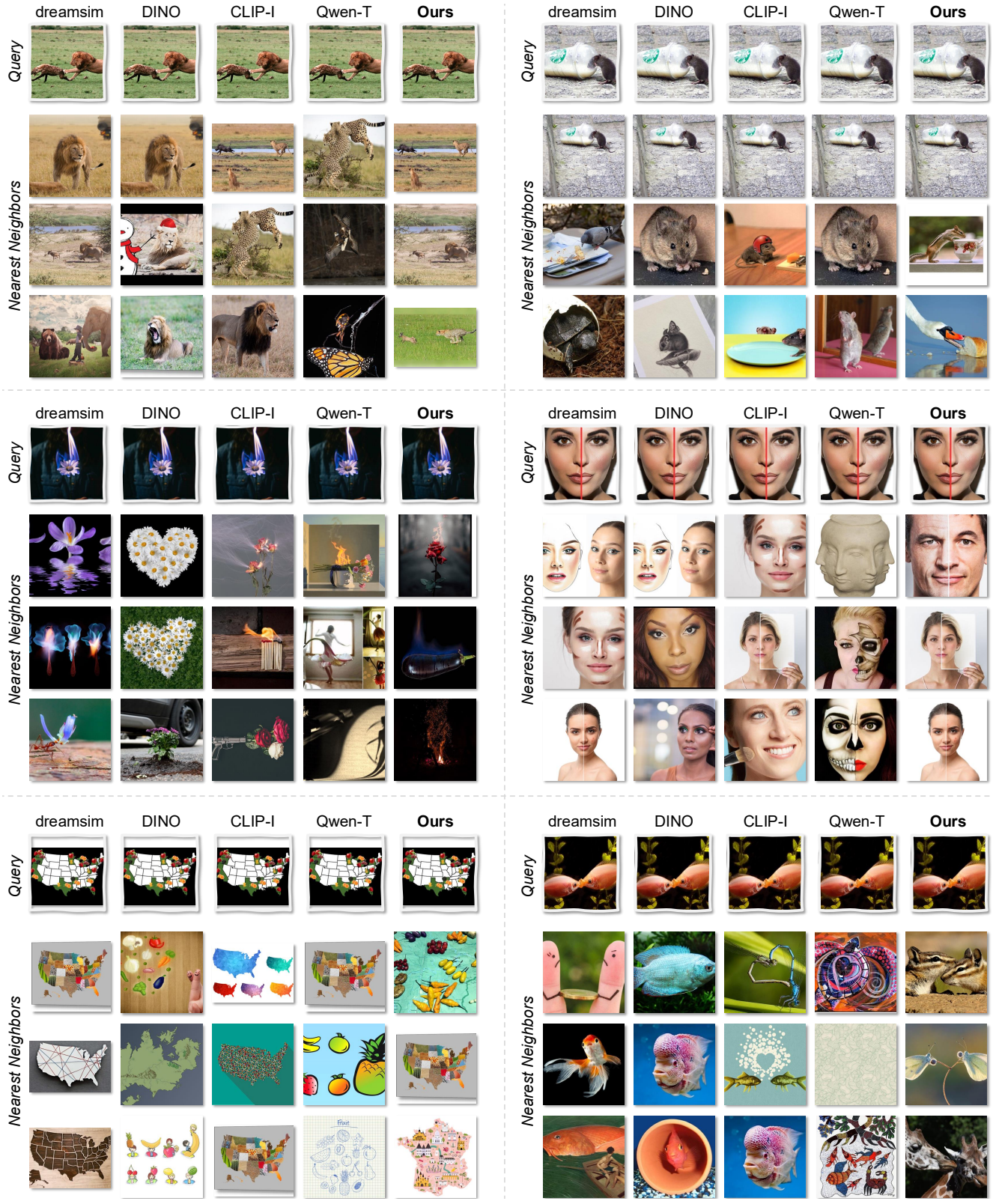

Figure 14. Additional results for image retrieval (1).

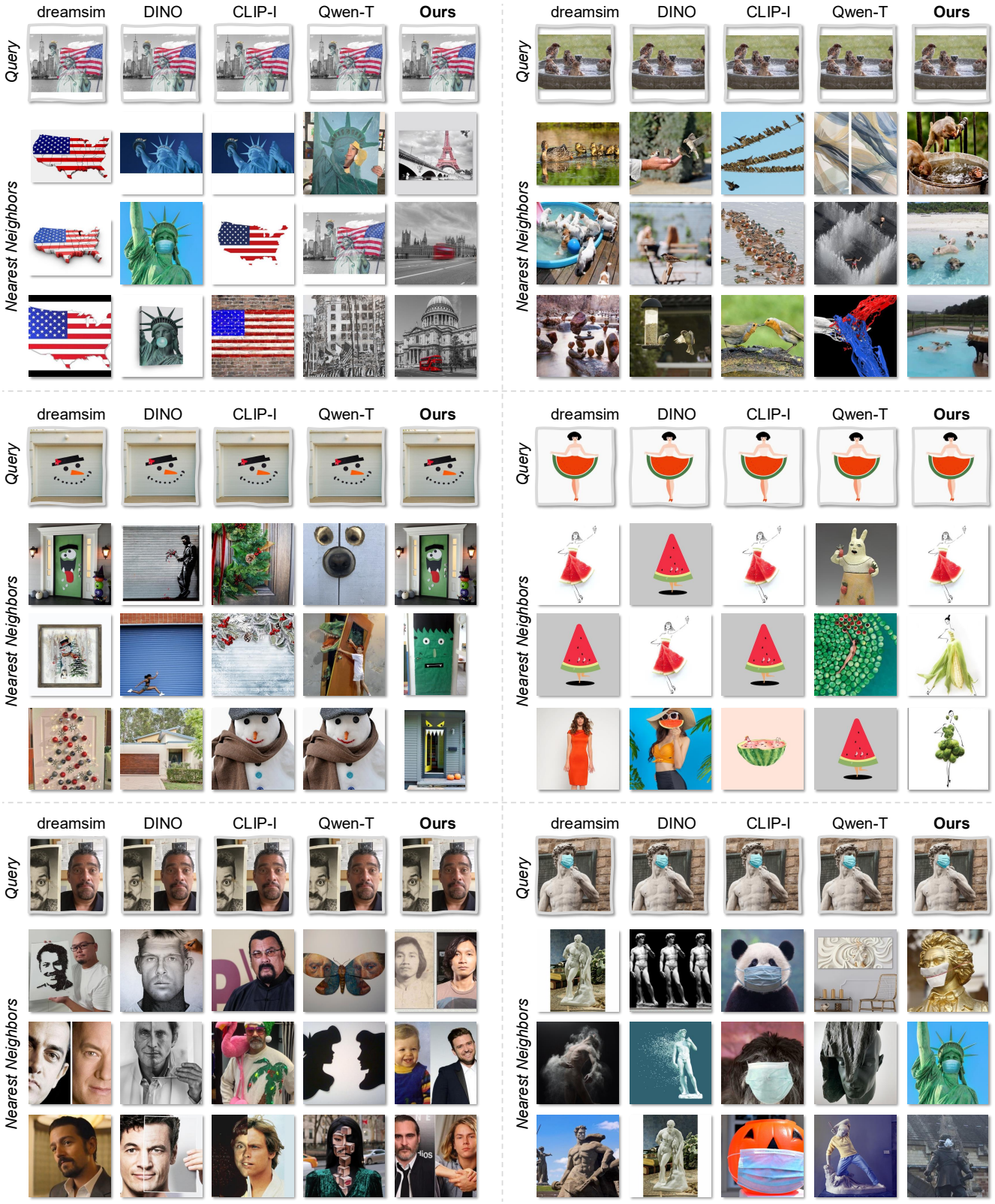

Figure 15. Additional results for image retrieval (2).