

Table 11: Evaluation criteria of the richness of text from face domain aspect.

1-Irrelevant: The text’s semantics are completely unrelated to human faces. The text does not mention any aspects of human faces or human body.
2-Minimal: The text contains a small amount of information or keywords related to human faces. The text briefly mentions a person or a face without providing substantial details.
3-Contextual: The text includes a relatively complete description, but not entirely focused on the face or body. The text provides some context or additional information beyond just the face or body.
4-Specific: The text contains a fairly specific description of the human face. The text describes the facial features, expressions, or accessories in some detail.
5-Comprehensive: The text includes a detailed description of the face and its attributes. The text provides a comprehensive description of the facial features, expressions, accessories, and additional context about the person.

Table 12: Evaluation criteria of the relevance between text and image from face domain aspect.

1-Unrelated: The text does not describe any elements present in the image. The image does not contain a human face, or the face described in the text is not present in the image.
2-Weakly Related: The text and the image share minimal connection. The text briefly mentions a person or a face, and the image contains a human face, but the description does not match the specific individual or facial features in the image.
3-Moderately Related: The text provides a description that partially matches the face in the image, but some key details are missing or inconsistent.
4-Strongly Related: The text and the image are closely connected. The text provides a specific description that matches the facial features, expressions, and accessories of the individual in the image.
5-Comprehensively Matched: The text provides a comprehensive and detailed description that matches all aspects of the face in the image, including facial features, expressions, accessories, and any relevant context.

Table 10: Training hyperparameters for FLIP models.

Hyper-parameter	Post-training	Scratch
batch size	1760/876/400	876
data size	80M	10M
epochs	3	5
learning rate	2e-5	3e-4
warming up (%)	1	5
scheduler	constant	cosine
augmentation	RandomCrop	same

A LLM DENOISING PROMPT

For text denoising, we instruct LLM with the prompt shown in Figure 8.

Few-shot examples:

[INPUT] Happy beautiful mother embracing her adorable baby. Family concept. Studio shot. - stock photo
[OUTPUT] Happy beautiful mother embracing her adorable baby. Family concept. Studio shot.

[INPUT] What to wear to a holiday party - Ashley Brooke www.ashleybrookedesigns.com
[OUTPUT] What to wear to a holiday party - Ashley Brooke

System Prompt:

Please denoise the input text according to the above examples that delete the part of the text that does not contain actual semantics. Directly output the result.

User Input:

#EF13594 #EF12868 2012 summer woman Fashion woman bronzing geometric prints sleeveless T-shirt top #EF24107 #EF12015

LLM Output:

2012 summer woman Fashion woman bronzing geometric prints sleeveless T-shirt top

Figure 8: Prompting LLM with few-shot examples to denoise a given text.

B FLIP TRAINING DETAILS

Our models are all trained on 8 Nvidia A100 80G GPUs. We use an open-source framework TencentPretrain [61], which has a consistent model implementation with the OpenAI CLIP model. The detailed hyper-parameters for post-training on CLIP and training from scratch are provided in Table 10.

C HUMAN EVALUATION DETAILS

We adopt human evaluation on the dataset quality (image-text relevance and richness). We design annotation criteria that divide the text richness and image-text relevance into 5 levels from the face domain aspect. We invited 5 volunteers to participate in manual evaluation. They first learn the annotation criteria from the Table 11 and Table 12, and then each evaluate 100 samples randomly sampled from the dataset according to the criteria.