

Role-SynthCLIP: A Role-Play Driven Diverse Synthetic Data Approach

Supplementary Material

7. Expert Role Generation Prompts

Inspired by LLM-Discussion [23], we use GPT-5 to automatically generate the roles used in our framework. These roles are not intended to be optimal; rather, they are preliminary designs to demonstrate the effectiveness of our approach. We also acknowledge that defining a broader spectrum of roles would further enhance data diversity, which we will explore in future work. The detailed prompt is shown in Table 6.

8. Expert Captioner Observation Prompts

We use the *agent role*, *agent speciality*, and *agent role prompt* to define distinct character profiles, enabling the MLLM to generate captions aligned with its assigned role. The detailed prompt is presented in Table 7 and we replace the final response with our local model name.

The sampling parameters we use are shown in Table 8. The parameters are configured following the official recommendations of Qwen.

9. Text-Image Pair Filter Prompt

Captions generated by multimodal large language models (MLLMs) may contain two types of errors: (i) inconsistencies between image and text, and (ii) deviation from the assigned role.

To address this, we leverage the teacher model GPT-5 to obtain training data containing both scores and reasoning paths. Using this data, we train a smaller and more cost-efficient filtering model to score image-text pairs. The filtering prompts used are shown in Table 9.

10. Filter Algorithm

Following SynthCLIP [15], image diversity is a key factor influencing model performance. While maintaining the quality of image-text pairs, we prioritize maximizing image diversity. After filtering, our dataset contains 879,406 unique images, each associated with an average of 3.81 distinct captions, reflecting both visual richness and linguistic variability. The pseudocode of our filtering algorithm is provided in Algorithm 1.

11. Training Hyperparameters

For a fair comparison, we adopt the same training parameters as Long-CLIP. Due to limited computational resources, we apply gradient accumulation to compensate for the smaller GPU count and maintain the same global batch

Algorithm 1 Role-Aware Cap-and-Refill Filtering Algorithm

Require: Dataset D with fields (image, caption, role, score)

Require: Hyperparameters K_{\max} , K_{\min} , TARGET_PAIRS

- 1: **Prefilter:** For each role r , keep the top 50% samples ranked by score.
 - 2: **if** dedup = True **then**
 - 3: Remove duplicates based on (image, caption).
 - 4: **end if**

 - 5: **Role-wise Normalization:**
 - 6: $s_i \leftarrow \frac{x_i - \mu_r}{\sigma_r + \varepsilon}$ for each role r

 - 7: **Global Sorting:** Sort all rows by normalized score s_i in descending order.

 - 8: **Initial Selection:**
 - 9: For each image, keep up to K_{\max} highest-scoring captions.
 - 10: Let K be the current kept set.

 - 11: **if** $|K| > \text{TARGET_PAIRS}$ **then** ▷ Over budget pruning
 - 12: Keep at least K_{\min} per image.
 - 13: Fill remaining quota by descending score order.
 - 14: **return** K
 - 15: **else if** $|K| < \text{TARGET_PAIRS}$ **then** ▷ Under budget refill
 - 16: Add top-1 caption for uncovered images.
 - 17: Refill remaining quota from candidate pool by score, respecting K_{\max} .
 - 18: **end if**

 - 19: **return** Final filtered dataset K .
-

size as Long-CLIP. The detailed training settings are summarized in Table 10.

Round 1 - User Prompt

What are some roles that are typically used in boosting precise image description, like different personas to precisely describe the image in different view.

Round 1 - GPT-5 Response

That's a really interesting question! When people try to boost the precision of image descriptions, they often use different roles or personas to look at the picture from complementary perspectives. This technique is used in annotation, accessibility work, and in multimodal AI prompting to generate richer and more precise outputs.

Here are some typical roles that can help:

1. The Observer of Details

Focuses on the specific visual attributes of an image such as objects, colors, textures, shapes, and lighting conditions. Strong emphasis on capturing factual, observable details.

2. Interpreter of Context

Identifies the possible meaning of visual elements, interpreting human expressions, implied actions, cultural references, and the situational context of the image.

3. Compositional Analyst

Examines the overall structure of the image, including arrangement of subjects, balance, perspective, depth, and spatial relationships. Highlights the visual organization and framing.

4. Narrative or Scene Setter

Synthesizes observed details and context to suggest a narrative, identifying the setting, possible actions, implied storylines, and the overall communicative intent of the image.

5. Emotional or Aesthetic Responder

Focuses on the images emotional resonance and aesthetic qualities, capturing mood, atmosphere, symbolic associations, and the subjective impression it may evoke in viewers.

The idea is that by combining multiple roles, you get a layered and more accurate description moving from raw perception to interpretation, while recognizing where assumptions end and evidence begins.

Round 2 - User Prompt

Great, now follow this format and generate a .json file for these roles: Remain type: openai, model name: gpt-3.5-turbo, the same. and the agent name = GPT Agent 1 - agent role

```
1 [
2   {
3     "type": "openai",
4     "model_name": "gpt-3.5-turbo",
5     "agent_name": "GPT Agent 1 - White Hat",
6     "agent_role": "White Hat",
7     "agent_speciality": "Information Analysis and Facts",
8     "agent_role_prompt": "Focuses on available data and past information, analyzing
9       trends and gaps in knowledge, striving for an objective viewpoint."
10  },
11  {
12    "type": "openai",
13    "model_name": "gpt-3.5-turbo",
14    "agent_name": "GPT Agent 2 - Red Hat",
15    "agent_role": "Red Hat",
16    "agent_speciality": "Emotions and Feelings Interpretation",
17    "agent_role_prompt": "Listens to and validates the emotional responses of the group
18      , understanding the values and intuition behind reactions, without judgment or
19      justification."
20  }
21 ]
```

Round 2 - GPT-5 Response

Below is the .json format including all the personas you specified:

```
1 [
2   {
3     "type": "openai",
4     "model_name": "gpt-3.5-turbo",
5     "agent_name": "GPT Agent 1 - Observer of Details",
6     "agent_role": "Observer of Details",
7     "agent_speciality": "Micro-level visual recognition",
8     "agent_role_prompt": "Focuses on the specific visual attributes of an image such as
          objects, colors, textures, shapes, and lighting conditions. Strong emphasis on
          capturing factual, observable details."
9   },
10  {
11    "type": "openai",
12    "model_name": "gpt-3.5-turbo",
13    "agent_name": "GPT Agent 2 - Interpreter of Context",
14    "agent_role": "Interpreter of Context",
15    "agent_speciality": "Semantic and situational interpretation",
16    "agent_role_prompt": "Identifies the possible meaning of visual elements,
          interpreting human expressions, implied actions, cultural references, and the
          situational context of the image."
17  },
18  {
19    "type": "openai",
20    "model_name": "gpt-3.5-turbo",
21    "agent_name": "GPT Agent 3 - Compositional Analyst",
22    "agent_role": "Compositional Analyst",
23    "agent_speciality": "Macro-level composition analysis",
24    "agent_role_prompt": "Examines the overall structure of the image, including
          arrangement of subjects, balance, perspective, depth, and spatial relationships
          . Highlights the visual organization and framing."
25  },
26  {
27    "type": "openai",
28    "model_name": "gpt-3.5-turbo",
29    "agent_name": "GPT Agent 4 - Narrative Setter",
30    "agent_role": "Narrative or Scene Setter",
31    "agent_speciality": "Story and situational framing",
32    "agent_role_prompt": "Synthesizes observed details and context to suggest a
          narrative, identifying the setting, possible actions, implied storylines, and
          the overall communicative intent of the image."
33  },
34  {
35    "type": "openai",
36    "model_name": "gpt-3.5-turbo",
37    "agent_name": "GPT Agent 5 - Emotional Responder",
38    "agent_role": "Emotional/Aesthetic Responder",
39    "agent_speciality": "Mood, tone, and subjective experience",
40    "agent_role_prompt": "Focuses on the image's emotional resonance and aesthetic
          qualities, capturing mood, atmosphere, symbolic associations, and the
          subjective impression it may evoke in viewers."
41  }
42 ]
```

Table 6. Prompts used for role generation.

Prompt for VLM to generate captions

You are a {agent role} whose specialty is {agent speciality}.
{agent role prompt}.
You will create an descriptions (max 150 words)/(max 30 words for short caption) for the given image, from the perspective of an {agent role}, highlighting {agent speciality}.
Output only the descriptions with no extra explanations. Do NOT use any title words such as: "Mood:", "Tune:", "Joyful simplicity:", "Muted Precision:", "Elegant simplicity:", "Muted Precision: "

Table 7. Prompt used for caption generation.

Parameter	Value
top_k	20
top_p	0.95
temperature	0.6
repetition_penalty	1.0

Table 8. Hyperparameter settings for generation.

Prompt for VLM to score an image-text pair

Here is Text Caption from a {agent role}, which role is {agent role prompt}
Text Caption: {caption}
Please evaluate if the provided text caption accurately represents the main features and objects of the image. The caption doesn't need to detail every aspect of the image, but it should capture its primary theme. Rate the overall quality of the text caption's match to the image on a scale of 1-100, considering the criteria mentioned. A higher score indicates higher level of image text matching.
Ensure that your scoring is nuanced and uses the entire range from 1 to 100, reflecting the subtle differences. The score should be given as an integer, with each number between 0 and 100 considered as a potential score, avoiding the tendency to round to multiples of 5 or 10. Please first output a single line containing the value indicating the scores. You must consider the role of the captioner when evaluating the caption. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias.

Table 9. Prompt for scoring image-text alignment and role consistency.

Configuration	Role-SynthCLIP Training
Global batch size	2048
Training epochs	6
Learning rate	1e-6
Warm-up steps	200
Learning rate scheduler	cosine
Optimizer	AdamW [22]
Optimizer hyperparameters	$\beta_1, \beta_2, \epsilon = 0.9, 0.999, 1e-8$
Weight decay	1e-2

Table 10. Training hyperparameters for Role-SynthCLIP.