Aligning Multi-Character Narrative Image Generation with Multi-Aspect Human Preferences

Supplementary Material

1. Implementation Details

In NIReward training, we set learning rate as 1e-4 and LoRA with 64 is employed. We set γ=0.7, and the model is trained on 1 epoch in each dimension. Our preference training procedure employed LoRA with the Adafactor optimizer, with a batch size of 1 pair and gradient accumulation over 2 steps. A learning rate of 1e-5 is used with 25% linear warmup. All training is conducted at a fixed square resolution of 1024×1024 for approximately 2,000 steps, with hyperparameters β, η, k, and b set to 5000, 0.3, 2, and 0.4, respectively.

2. ADPO algorithm

Algorithm 1 ADPO Algorithm

Dataset: Prompt set Y, Reference set C, instructions for reward model $\mathcal{Q} = \{q_1, q_2, \ldots, q_K\}$.
Input: Diffusion model p_θ, reward model r, the number of generated images per prompt N, reject threshold th.
Initialization: Preference set $\mathcal{D} \leftarrow \emptyset$.
for each $y \in Y$ **do**
  Random sample $c \in C$.
  Sample a set of images $\{x_0^1, x_0^2, \ldots, x_0^N\} \sim p_\theta(x_0|y, c)$.
  for $i = 1$ to N **do**
    Reward set $R_i \leftarrow \emptyset$.
    for $k = 1$ to K **do**
      Add reward $r(x_0^i, y, c, q_k)$ to R_i
    end for
  end for
  for $i, j \sim \{1, 2, \ldots, N\}$ **do**
    if x_0^i dominates x_0^j and $r(x_0^i, y, c, q_k) \geq th$ $(\forall k \in \{1, \ldots, K\})$ **then**
      $a = \frac{1}{K}\sum_{k=1}^{K}(r(x_0^i, y, c, q_k) - r(x_0^j, y, c, q_k))$
      Add $\{x_0^i, x_0^j, y, c, a\}$ to $\mathcal{D}$.
    end if
  end for
end for
for $\{x_0^i, x_0^j, y, c, a\} \sim \mathcal{D}$ **do**
  Update the gradient p_θ by Eq.9.
end for

3. Examples of NIReward

Fig. 1 shows examples where specific models(CLIP, ArcFace, Aesthetic Score) fail to correctly identify preferences, whereas NIReward correctly infers the preferences through critical reasoning. For instance, in the first case, NIReward identifies the incompleteness of details such as the background in the losing sample.

4. Limitations and Future Works

In this section, we discuss some limitations and future work.
Annotation scale and quality. The larger scale of annotation datasets is still needed for better reward model training. In the future, we plan to further increase the volume of human preference data to enhance model performance. Additionally, cross-validation by human experts will be implemented to improve dataset quality and ensure the reliability of the collected preferences.
Comparison with RL-based methods. In this paper, we only discuss DPO-based approaches and have achieved significant improvements. In future work, we will conduct comparative studies with RL-based methods such as DDPO and further refine our method based on these comparisons. This will provide a more comprehensive understanding of the relative strengths and limitations of different preference optimization techniques.

5. Broader Impacts

Our research on aligning multi-character narrative image generation with human preferences can significantly enhance human creative capabilities through generative AI systems, particularly in domains such as comic creation and film production. The technologies we have developed have the potential to lower entry barriers across various visual narrative fields including comic design, storyboarding, and visual content creation, thereby facilitating a more diverse range of voices in media production. Nevertheless, we acknowledge that advancements in identity consistency for generated images could potentially be misappropriated for creating misleading or deceptive content. To address this concern, we emphasize the critical importance of implementing robust watermarking and detection systems, complemented by the promotion of responsible usage guidelines and ethical frameworks. This balanced approach aims to maximize the creative benefits of our technology while mitigating potential societal risks.

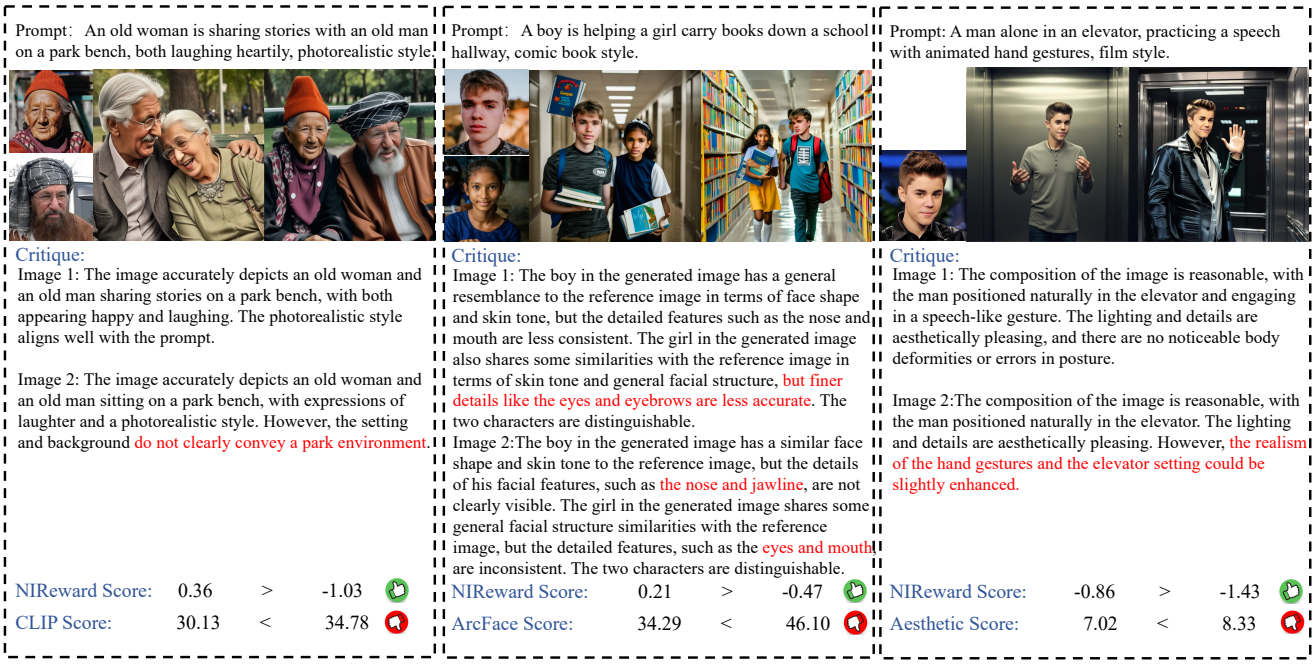


Figure 1. Comparison results of NIReward and specific models.

6. More details of NI-RLHF Dataset

6.1. Statistics of NI-RLHF

Fig 2 presents the statistical analysis of NI-RLHF. The proportion of images generated by PuLID and OMG is relatively small due to identity blending issues in PuLID and the unstable identity consistency inherent in OMG's training-free methodology. Regarding character distribution within the dataset, man and woman characters, being the most commonly utilized, constitute a higher percentage. The distribution of narrative prompt lengths is illustrated in Fig. 2(c), demonstrating that the dataset encompasses both complex descriptions with extended text and simpler narrative scenarios with concise prompts, thereby facilitating comprehensive evaluation.

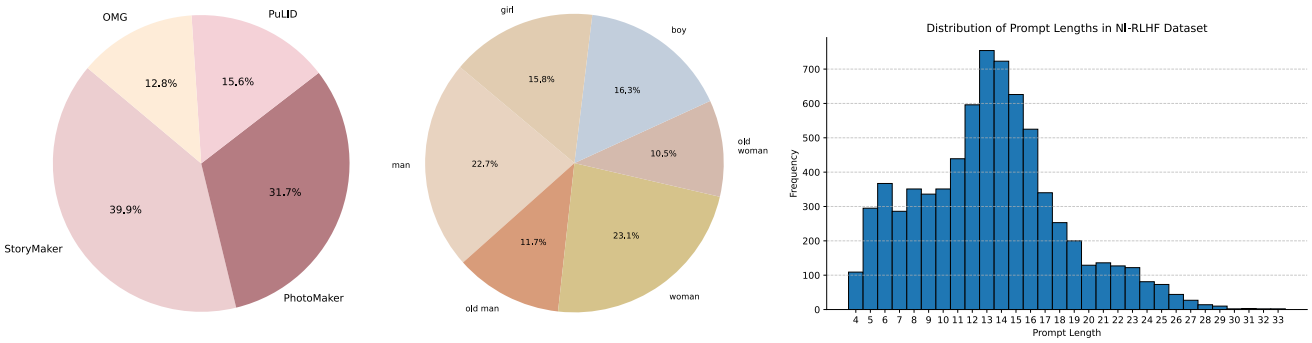
6.2. Prompt Generation Template

We begin our NI-RLHF Dataset by generating a diverse set of prompts. We establish six character categories, including man, woman, elderly man, elderly woman, boy, and girl. We use GPT4o, deepseek, Kimi 1.5, Gemini 1.5 flash, Qwen2.5 to generate narrative prompts, by repeatedly using the following prompt:

Narrative Prompt Generation Instruction Template

You are a prompt generator, please give me 100 narrative prompts.
The prompt should be a description of a single-frame scene from a comic or movie, containing one or two characters, which can be two of [man, woman, boy, girl, old man, old woman]. The characters should not be in the same class, e.g., should not have "two men" or "man1 and man2". The description of the character may include expressions and actions, but should not include identifying features such as appearance or clothing. Characters can interact with each other or perform actions independently.
The prompt should be reasonable, clear, concise, and diverse. Example: 1. A boy stopped on the city sidewalk, looking up at the sky. 2. A man and a woman are playing a baseball video game. 3. The woman is feeding ducks while the old man watches silently. 4. A woman and a man are hugging on the square. 5. In a dimly lit alley, a man shoots at a woman.
Please imagine and generate.

We then randomly select a style from realistic, film, photorealistic, anime sketch, comic book, or no style specification to incorporate into the prompt. This approach accom-



(a)Distribution of Models (b)Distribution of Characters (c)Distribution of Prompt Lengths

Figure 2. Statistics of NI-RLHF.

modates various application scenarios such as comic generation and film generation. Additionally, to enhance prompt diversity, we extract action descriptions from the COCO and Kinetics datasets to stimulate more varied outputs from the LLMs.

6.3. Annotation Guidelines

We design the scoring and preference annotation instructions to enable the MLLM to assist in annotating the generated images.

Score Annotation Instruction for Prompt Following

Task Definition
You will be provided with an image and text prompt. Text prompt involves 1-2 characters. As an experienced evaluator, your task is to analyse and evaluate the prompt-following quality between the image and the text prompt. You first need to summarize the prompt-following quality **critique** based on the following scoring criteria in 1-2 short sentences. Then you need to score the prompt-following quality.

Scoring Criteria
1. Character accuracy: Does the number, gender, and age of characters match the prompt accurately?
2. Action accuracy: Do the characters' actions in the image match the prompt accurately?
3. Expression accuracy: Do the characters' expressions in the image match the prompt accurately?
4. Detail completeness: Are the other elements in the image consistent with the prompt, such as the background?

Scoring Range
Based on these criteria, a specific score from 1.0 to 5.0(can be a decimal) can be assigned to determine the level of semantic consistency:
- Bad (1): No correlation. Character inaccurate. The image does not reflect any of the key points or details of the text.
- Poor (2): Weak correlation. Actions inaccurate. The image addresses the text in a very general sense but misses most details and nuances.
- Normal (3): Moderate correlation. Expressions inaccurate. The image represents the text to an extent but lacks several important details or contains some inaccuracies.
- Good (4): Strong correlation. Detail incompleteness. The image accurately depicts most of the information from the text with only minor omissions or inaccuracies.
- Excellent (5): Near-perfect correlation. The image captures the text's content with high precision and detail, leaving out no significant information.
For example, Score: 3.4(Normal)

Input Format
Every time, you will receive a text prompt and an image.
Please carefully review the image and text prompt. And then analyse and evaluate the semantic consistency.

Output Format
Reason: [Your critique]
Score: [Your Score]

Text prompt
{Prompt}
Image
{Generated Image}

Score Annotation Instruction for Identity Consistency

Task Definition
You will be provided with a generated image, 1-2 reference identity images representing different characters, and a text prompt. The first image is generated based on the identity images provided as references by the model. The text prompt involves 1-2 characters. As an experienced evaluator, your task is to analyse and evaluate the identity consistency between the image and the reference identity images. Please carefully review the image, and then you need to provide a concise critique of 1-2 sentences based on the following criteria, and finally score the identity consistency.

Scoring Criteria
1. Overall facial identity consistency: Are the general identity features of each person, such as face shape, facial feature proportions, and skin tone etc., consistent with the reference image?
2. Detail-level facial identity consistency: Are the detailed identity features of each person (such as eyes, eyebrows, mouth, beard, etc.) consistent with the reference image?
3. Identity distinguishability: Are the identity features between different characters clearly distinguishable? If the prompt involves only one character, skip this question.
Note: The consistency of identity is not evaluated based on clothing or hair, but only focuses primarily on facial features.

Scoring Range
Based on these criteria, a specific score from 1.0 to 5.0(can be a decimal) can be assigned to determine the level of semantic consistency:
- Bad (1): Identity is highly inconsistent. Identity distinguishability is bad when there are multiple characters. Each character does not resemble the reference image.
- Poor (2): Weak consistency, with high overall identity inconsistency. The image addresses the text in a very general sense but misses most details and nuances.
- Normal (3): Moderate consistency, with high detail-level identity inconsistency. The characters represent identity consistency to some extent, but lack several identity details or contain some inconsistencies.
- Good (4): Strong consistency, with low detail-level identity inconsistency. The image has only a few details that do not match the identity characteristics.
- Excellent (5): Near-perfect consistency. Identity characteristics are highly consistent, and identity distinction is clear.
For example, Score: 3.4(Normal)

Input Format
Every time you receive a text prompt, a generated image, and 1-2 reference images. Please carefully review the images and text prompt. And then analyse and evaluate the identity consistency.

Output Format
Reason: [Your critique]
Score: [Your Score]

Text prompt
{Prompt}

Images
{Generated Image}
{Reference Images}

Score Annotation Instruction for Visual Quality

Task Definition
You will be provided with an image and text prompt. Text prompt involves 1-2 characters. As an experienced evaluator, your task is to analyse and evaluate the visual quality of the image, which concentrates on the quality of images, and especially whether objects in generated images are realistic, aesthetically pleasing, and with no errors in the image itself. Please carefully review the image, and then you need to provide a concise critique of 1-2 sentences based on the following criteria, and finally score the visual quality.

Scoring Criteria
1. Reasonable Composition: The framing of the scene should adhere to the logic of a real-world setting, including the positioning of characters relative to one another and the proportion of figures to the scenery.
2. Fidelity: The image should be aesthetically pleasing and realistic.
3. No Body Deformity problem: Characters should not exhibit any physical posture deformities, such as distorted hands or other bodily anomalies.

Scoring Range
Based on these criteria, a specific score from 1.0 to 5.0(can be a decimal) can be assigned to determine the level of semantic consistency:
Bad (1): The visual quality is very poor. The composition of the image is unreasonable, does not meet aesthetic standards, and has issues with body deformity.
Poor (2): There are noticeable body deformity issues and compositional flaws.
Normal (3): The composition is reasonable, with no body deformity issues, but it does not meet aesthetic standards.
Good (4): The composition is reasonable, with no body deformity issues, but lacks realism in details.
Excellent (5): The aesthetics and fidelity are very good.
For example, Score: 3.4(Normal)

Input format
Every time you will receive a text prompt and an image. Please carefully review image and text prompt. And then analyse ans evaluate the semantic consistency.

Output Format
Reason: [Your critique]
Score: [Your Score]

Text prompt
{Prompt}

Images
{Generated Image}

Preference Annotation Instruction for Prompt Following

Task Definition
You will be provided with two images and a text prompt. Text prompt involves 1-2 characters. As an experienced evaluator, your task is to evaluate which image has better prompt-following quality with the text prompt and provide reasons. You need to first evaluate the following criteria.

Scoring Criteria
1. Character accuracy: Does the number, gender, and age of characters match the prompt accurately?
2. Action accuracy: Do the characters' actions in the image match the prompt accurately?
3. Expression accuracy: Do the characters' expressions in the image match the prompt accurately?
4. Detail completeness: Are the other elements in the image consistent with the prompt, such as the background?

Input format
Every time, you will receive a text prompt and two images.
Please carefully review the images and text prompt. And then analyse and evaluate the prompt-following.

Output Format
Image 1/2 is better. [Reason]

Text prompt
{Prompt}

Images
{Generated Image1}
{Generated Image2}