Let's imagine a quirky, pint-sized cartoon monster with bee vibes. Colorful and detailed, but no pink!

Prompt: Fantz cartoon monster|bee design, ultra detailed, random colors, masterpiece, best quality, detailed face, two eyes, no pink

Model: Small_Mostersv2.
A Lora Model that generates funny monsters. It is recommended to use it with the Realistic Vision model for better results.

```
Arguments:
{'sampler': 'DPM++ 2S a Karras',
  'steps': 42,
  'cfg_scale': 4,
  'width': '600',
  'height': '600'
}
```

Let's create a lifelike photo of a charming 72-year-old German hippie man outdoors.

Prompt: RAW photo, 8k UHD, high detailed skin, handsome german hippie, 72yo, braid, outdoors, gorgeous 8k uhd, dslr, soft lighting, high quality, film grain, Fujifilm XT3

Model: devlishphotorealism_v40.
This packet is suitable for creating realistic and detailed images with a focus on photo-like quality.

```
Arguments:
{'sampler': 'DPM++ SDE Karras',
  'steps': 60,
  'cfg_scale': 7,
  'width': '512',
  'height': '512'
}
```

Well, picture this: a dashing gent, uh, with rugged refinement. You know, fair-skinned, boy-next-door vibe, kinda cinematic. Blazer, slim trousers, tie.

Prompt: realistic, masterpiece, best quality, detailed, cinematic, dynamic lighting, soft shadow, detailed background, professional photography, detailed face, realistic hair, realistic eyes, muscular, manly, photo of a handsome man, sexyloincloth, jacket, white shirt, necktie, collared shirt, slim waist, thick thighs, medium hair, brown hair, sitting

Model: hellorealisticman_V12a.
A packet that focuses on creating realistic images of male characters. It is suitable for creating detailed and high-quality images of handsome and muscular men in various settings.

```
Arguments:
{'sampler': 'DPM++ 2M Karras',
  'steps': 28,
  'cfg_scale': 7,
  'width': '512',
  'height': '768'
}
```

Imagine Miku, aqua-haired beauty in white, gazing at us with sparkling eyes behind stylish glasses.

Prompt: masterpiece, best quality, 1girl, solo, hatsune miku, white outfit, glasses, aqua hair, looking at viewer sparkling eyes, light particles, pastel colors, detailed eyes, high resolution, backlighting, ultra detailed, illustration, screen space, solo, standing, upper body, viewer, wallpaper,Grey background, hatsune miku, 8k

Model: rolua.
Rolua style LoRA is suitable for creating highly detailed and beautiful anime and manga style images. It is recommended to use this style with anime models

```
Arguments:
{'sampler': 'Euler a',
  'steps': 20,
  'cfg_scale': 9,
  'width': '896',
  'height': '1344'
}
```

Yo, I'm aiming for a super realistic, movie-quality 4K portrait of Ellie from 'The Last of Us'. Gotta look like a real photo!

Prompt: the last of us, Ellie, natural hair, realistic portrait, 4k, supreme detail, highly detailed, artstation, smooth, sharp focus, cinematic lighting

Model: epicdream_lullaby
This style is focused on achieving realistic and highly detailed images. It is suitable for anime, landscapes, photorealistic art, and portraits.

```
Arguments:
{'sampler': 'DPM++ 2M SDE Karras',
  'steps': 30,
  'cfg_scale': 7,
  'width': '512',
  'height': '768'
}
```

I've been dreaming of this Tolkien-esque elf woman. Picture Evangeline Lilly with a fantasy twist and cinematic vibes.

Prompt: SFW, Tolkien inspired setting, 1girl, solo, Evangeline Lilly|Ragnarok|Rhulon, green hair, white dress, magical pants, belt, beautiful,expressive, detailed, cinematic, soft light, backlighting, professional, 4k, max detail, highres, ...

Model: rawXs_v2
a packet that specializes in generating highly detailed photos with a realistic and photorealistic style.

```
Arguments:
{'sampler': 'DPM2 a Karras',
  'steps': 50,
  'cfg_scale': 7,
  'width': '416',
  'height': '608'
}
```

Figure 1. Examples of single inputs with step-wise outputs.

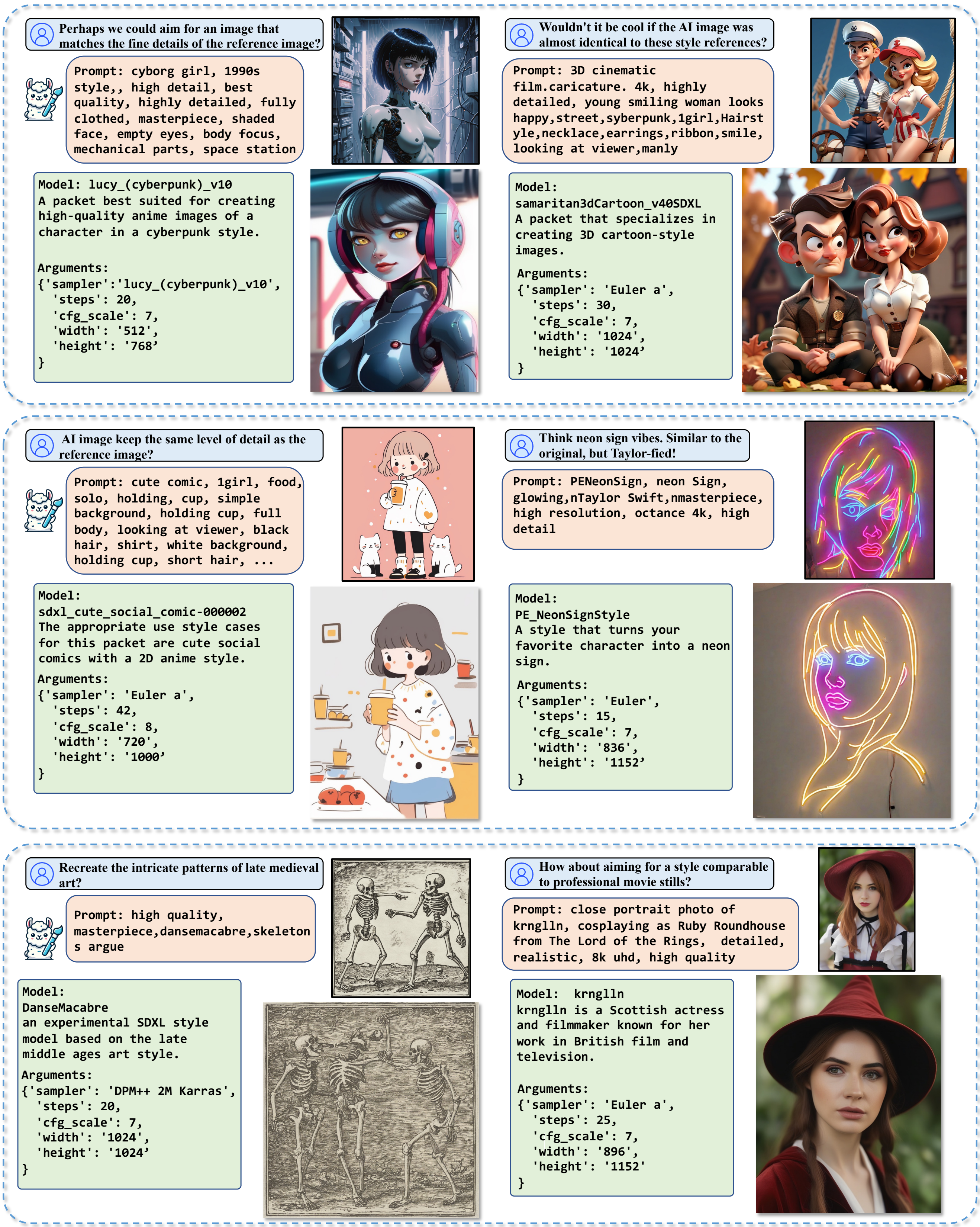

Figure 2. Examples of multimodal inputs with step-wise outputs. The image in the top-right corner represents the input reference image.

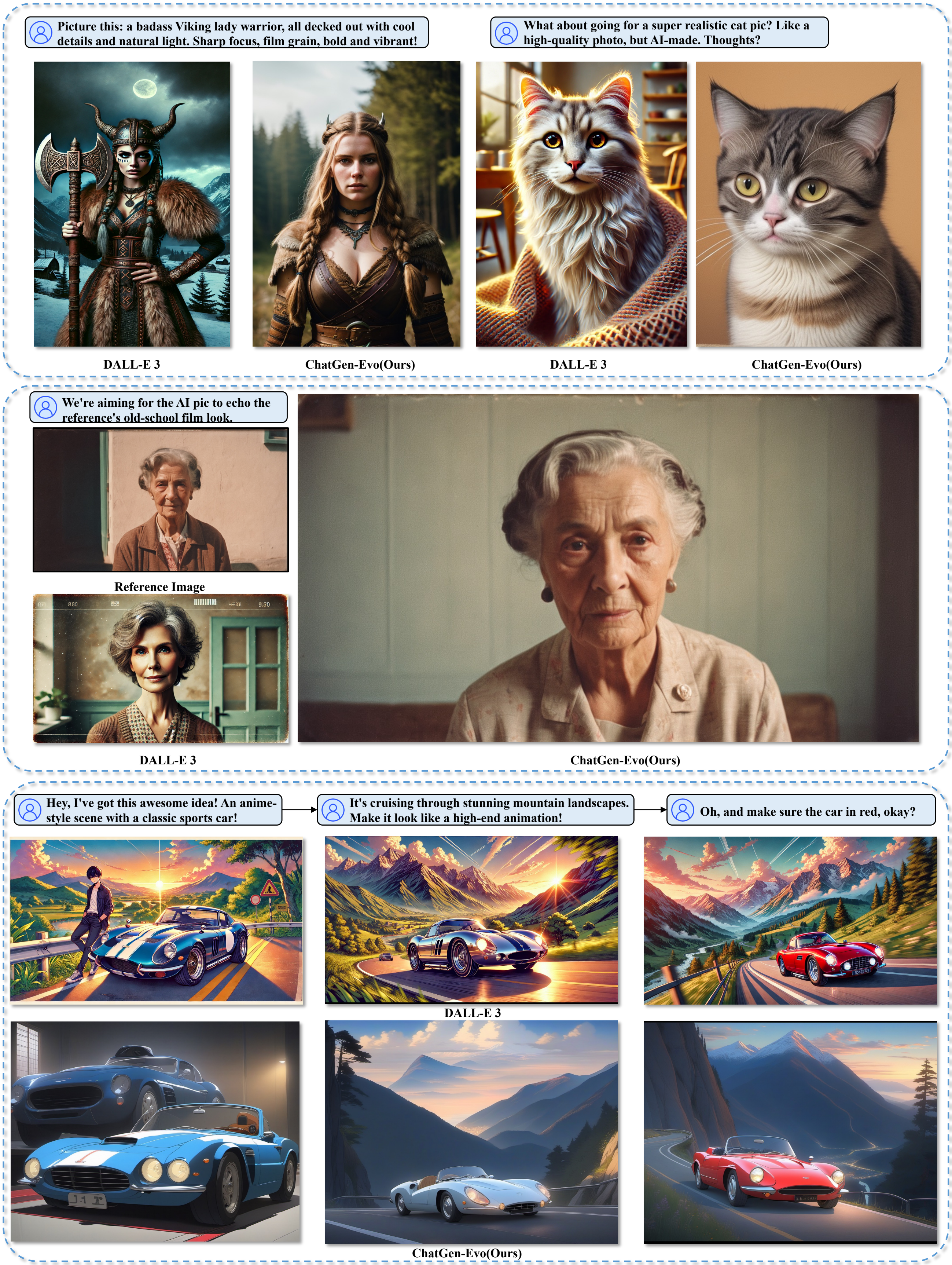

Figure 3. Examples of images generated by ChatGen-Evo and DALL-E 3. Three rows represent single, multi-modal and historical inputs, respectively.

ChatGen: Automatic Text-to-Image Generation From FreeStyle Chatting

Supplementary Material

This supplementary material offers extensive additional details and more qualitative and quantitative analysis complementing the main paper. The content is organized as follows:

- More visualization results (Appendix 1)
- 8B version of ChatGen-Evo (Appendix 2)
- More details of chatting generation (Appendix 3)
- More details of benchmark construction (Appendix 4)
- Visualization of model descriptions (Appendix 5)
- Potential limitations and challenges (Appendix 6)
- More details of training and inference (Appendix 7)
- More details of metrics for evaluation(Appendix 8)
- More details of human evaluation (Appendix 9)
- Prompts for chatting generation (Appendix 10)

1. More visualization results

1.1. More samples with step-wise outputs

In Figures 1 and 2, we present the step-wise outputs and final images of ChatGen-Evo. It can be observed that ChatGen-Evo effectively rewrites high-quality professional prompts based on the user's freestyle input. Furthermore, ChatGen-Evo selects suitable models to match the user's desired style or character. Finally, it generates appropriate argument configurations to ensure the high quality of the resulting images. These high-quality images, produced through well-designed step-wise outputs, demonstrate the value of Automatic T2I. It relieves users from tedious steps and automates the production of desired images directly from their freestyle input.

1.2. Comparisons with commercial models

In Figure 3, we compare the image quality of our method with the advanced commercial model DALL-E 3 [1]. While DALL-E is capable of generating high-quality images, its style is predominantly limited to a single type (anime-style). This limitation arises from its reliance on a single model, which cannot fully accommodate diverse and personalized styles. This highlights the value of our approach, which performs significantly better in scenarios requiring realistic styles or other personalized outputs. Additionally, we compare the results for history inputs. While DALL demonstrates strong performance, our open-source project contributes to advancing research in this field and helps bridge the gap between commercial models and open-source solutions in these aspects.

2. 8B version of ChatGen-Evo

In Table 2, we present the performance of the 8B version of ChatGen-Evo on ChatGenBench. The results indicate that while the 8B model brings some performance improvements compared to the 2B version, the gains are relatively minor. The primary bottleneck lies in the limited improvement in the prompt rewriting component. These findings suggest that merely increasing model size offers diminishing returns. Exploring more efficient strategies and enhancing the scale and quality of the prompt rewriting data is crucial for further advancing automatic T2I.

3. More details of chatting generation

To ensure the quality of generation, we employed multiple versions of (M)LLMs to complete all generation tasks, with all (M)LLMs displayed in Table 2. Additionally, we set the temperature parameter to 0.9 to make the model more stochastic during the generation. We also utilized BertScore to filter out results with a similarity greater than 0.8, ensuring diversity in the outputs.

4. More details of benchmark construction

We perform multiple filtering rounds to ensure the ChatGenBench's quality. The following steps are executed sequentially to ensure a thorough and systematic refinement of the dataset:

4.1. Length Filtering

We remove all results where input text exceeds 40 words and discard history data with more than six turns of dialogue to ensure the conciseness.

4.2. Colloquialism Check

To ensure that our dataset aligns with natural, everyday language, we employ the Spacy [4] library for data filtering. The process involves the following steps:

- **Colloquial Vocabulary Detection**: Identify the presence of common colloquial terms such as "wanna", "let's", "you know".
- **Colloquial Structures**: Analyze the use of pronouns and interjections to assess colloquial expressions.
- **Part-of-Speech (POS) Analysis**: Examine POS tags to exclude texts that are overly technical or lack conversational elements.

Table 1. The Step-wise and Final evaluation results of ChatGen-Evo (2B) and ChatGen-Evo (8B) on ChatGenBench.

		Step-wise Evaluation			Final Evaluation				
		Prompt BertScore ↑	Selection Acc ↑	Config Acc ↑	FID Score ↓	CLIP Score ↑	HPS v2 ↑	Image Reward ↑	Unified Metric ↑
Supervised	ChatGen-Evo (2B)	0.247	0.328	0.537	19.1	72.9	25.1	8.9	65.9
	ChatGen-Evo (8B)	0.269	0.422	0.577	18.5	73.1	25.3	14.4	68.0
Few-shot	ChatGen-Evo (2B)	0.283	0.231	0.493	40.7	72.5	25.0	5.1	59.2
	ChatGen-Evo (8B)	0.291	0.341	0.532	40.2	72.9	25.0	6.4	59.8

Table 2. Overview of MLLMs utilized in chatting generation. The table lists different versions of MLLMs on the left, with corresponding approximate numbers of generated chatting queries shown on the right.

MLLM	Numbers
gpt-4o-2024-08-06	≈100,000
gpt-4o-mini-2024-07-18	≈70,000
gpt-4-turbo-2024-04-09	≈70,000
claude-3-5-sonnet-20240620	≈100,000

- **Syntactic Complexity**: We analyze sentence structures to filter out texts with complex syntax that may not represent natural speech patterns.

4.3. LLM-Based Evaluation

To ensure our dataset aligns with a conversational tone, we utilize advanced LLM, GPT-4o, to evaluate and select the most suitable data. The LLM is configured to assess these texts, focusing on their colloquialism and alignment with natural, everyday language. Each text entry is assigned a score between 0 and 5, where 0 indicates a formal or unnatural tone, and 5 represents a highly conversational and natural tone. Entries scoring below 4 are excluded, ensuring that only those with a strong conversational tone are retained. This process results in a refined dataset that closely matches the desired chatting tone.

4.4. Manual Verification

The final step in the filtering process is manual verification, where phd-level reviewers carefully inspect the filtered dataset to identify and eliminate any inappropriate or irrelevant samples. This step is essential to ensure that the dataset maintains high quality and aligns with the desired conversational tone. During manual verification, reviewers analyze each text entry to confirm that it meets all criteria, including natural language use, appropriateness for the intended context, and relevance to the task. This meticulous manual review guarantees that the final benchmark is both accurate and suitable for the intended application.

5. Visualization of model descriptions

We present a word cloud generated from model descriptions, showcasing the diversity and importance of various attributes. This visualization highlights key themes such as realism, character design, artistic styles, and fashion, demonstrating the model's broad applicability.

6. Potential limitations and challenges

We identify two critical challenges that require further exploration to enhance the system's efficiency and reliability: (1) **Efficient High Workload Management:** Deploying multiple models simultaneously can increase latency in high-demand scenarios. It is crucial to optimize the system's infrastructure for better concurrency and enhance core algorithms to minimize delays, ensuring efficient handling of high workloads. (2) **Multi-Model Integration and Conflict Management:** The integration of multiple models introduces complexities such as conflicts and deployment challenges. Developing a robust model management framework that can adeptly handle model dependencies and resolve conflicts is essential.

7. More details of training and inference setups

Batch Size. To handle the large-scale training data, we set a global batch size of 128 across 8 GPUs, leveraging gradient accumulation to optimize memory usage and training stability. The batch size per GPU is 4.

Preprocess. Input images are resized to 448×448 pixels using bicubic interpolation. Standard image augmentation techniques such as random cropping, horizontal flipping, and color jittering are applied during training. Images are normalized using the ImageNet mean and standard deviation values.

Inference. During inference, the maximum output token length is set to 1,024. For outputs that do not conform to the fixed format, we employ scripts to automatically correct them, ensuring the accuracy of data used for text-to-image (T2I) tasks. Additionally, the baseline method may exhibit hallucinations, generating references to non-existent models. We detect such instances and assign default models and parameters to maintain consistency.

8. More details of metrics for image quality evaluation

We provide detailed explanations and implementations for each metric used in our image quality evaluation.

Fréchet Inception Distance (FID) [3] is a metric that quantifies the similarity between two datasets of images by calculating the Fréchet distance between two multivariate Gaussian distributions. To obtain these feature representations, both sets of images are processed through a pre-trained Inception-V3 [5], which extracts high-level features from the images. In our evaluation, FID is computed between automatically generated images and human-validated high-quality images. We employ the official code of FID[1] to compute this score.

CLIPScore [2] aims to evaluate the correlation between text and image using the cosine similarity of their respective embedding obtained through CLIP. CLIPScore is able to achieve a high correlation with human judgment for the text-image alignment. We employ the official code of CLIPScore[2] to compute this score.

Human Preference Score v2 (HPSv2) [6] is an advanced metric designed to evaluate text-to-image generative models by aligning their outputs with human preferences. HPSv2 addresses the limitations of traditional automated metrics by incorporating human judgment into the evaluation process, which is trained on HPD v2, a large-scale dataset comprising 798,090 human preference choices. We compute HPSv2 based on the official code[3].

ImageReward [7] metric is measured by a general-purpose text-to-image human preference reward model, which is trained on 137k pairs of expert comparisons. By evaluating generated images based on factors such as realism, prompt alignment, and aesthetic appeal, ImageReward provides a robust metric for aligning with human preferences. We compute it based on the official code[4].

Unified Metric aims to provide an intuitive and comprehensive measure of image quality. Following DiffAgent [8], we normalize and combine the above four metrics into an aggregated score. Each of above scores are normalized to the range [0,1] and the final score is computed by averaging the value:

$$\begin{aligned} S_{unified} =&\frac{1}{4}\,((1 - norm(S_{fid})) + norm(S_{clipscore}) \\ &+ norm(S_{hps}) + norm(S_{reward})). \end{aligned} \tag{1}$$

By integrating these diverse metrics, the Unified Metric offers a more nuanced and comprehensive understanding of image quality, facilitating more informed comparisons and evaluations of generated images.

9. More details of human evaluation

Sampling Process. The image pairs used for the human evaluation are sampled proportionally and randomly from the original benchmark. This process ensured a balanced and representative selection of image pairs for evaluation. Specifically, we generated 2,000 pairs for the supervised setting and 1,000 pairs for the few-shot setting, as mentioned in the main text.

Evaluation Process. We enlist five PhD-level volunteers to participate in the evaluation process. The assessment is conducted using an interactive interface, as shown in Figure 4. In this interface, two images generated from the same input are displayed side by side. One image is generated by ChatGen-Base(8B) and the other by ChatGen-Evo(2B). The left and right positions of the images were randomly assigned to avoid positional bias.

Volunteers are tasked with comparing the two images based on two criteria: image quality and relevance to the given input. For each criterion, they select the image they judge to be better. The system records their choices, and the results are aggregated to compute scores for both models.

Table 3. Summary of Sampling data for human evaluation.

	Total	Single	M-Modal	History
Benchmark	14,564	11,011	1,668	1,132
Supervised	10,240	8,009	1,099	1,132
Sampling data	2,000	1,600	200	200
Few-Shot	4,324	3,002	569	753
Sampling data	1,000	800	100	100

10. Prompts for chatting generation

In this section, we provide details of prompts for multimodal-input and history-input, as well as examples provided, to facilitate understanding of the chatting genera-

[1] the official code of FID: https://github.com/bioinf-jku/TTUR
[2] the official code of CLIPScore: https://github.com/jmhessel/clipscore
[3] the official code of HPSv2: https://github.com/tgxs002/HPSv2
[4] the official code of ImageReward: https://github.com/THUDM/ImageReward

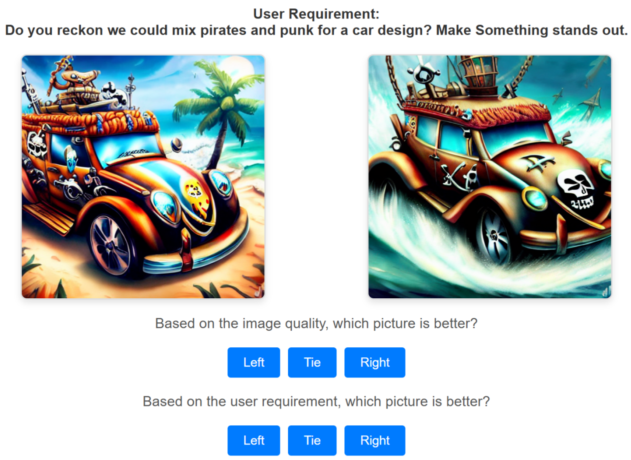

Figure 4. Screenshot of the user interface for the human evaluation.

tion process. Additional prompts, including more examples, will be provided when the project is open-sourced.

Prompt: Multimodal-Input chatting generation

System: You are a professional user experience designer who plays various personas to convert complex and professional content for non-professional users. Please merge the following prompt, along with any possible reference images and the description of the model, into a single freestyle query. Remove any obvious details that non-professional users would naturally want to avoid, making it similar to what non-professional users may write, so they can easily understand and use it. And the converted user-friendly prompt should be colloquial and as brief as possible. Moreover, combining the description of the model for generating reference images and the specific content of the image, focuses on describing how the generated image is similar in style to the reference images, such as light, texture, color, details. And slightly describe how important specific elements of the generated image should resemble the reference images, such as color palette, composition, or the way certain details are rendered.
Require: {ROLE}, {PROMPT}, {REFERENCE IMAGE} {MODEL}, {Example 1,...,n}
Input: You are playing the {ROLE}. Please generate a single text query based on the following professional {PROMPT} for the {MODEL} with the reference image {REFERENCE IMAGE}. You can refer to following examples: {Example 1,...,n}.

Prompt: History-Input chatting generation

System: You are a professional user experience designer who plays various personas to convert complex and professional content for non-professional users. Now, Please help me convert a text-to-image prompt into a natural dialogue context. Dialogue should be **colloquial and accessible**, similar to people's daily conversations. Dialogue should be broken down into short sentences and presented in the form of human and gpt dialogue. Dialogue should **be filled with rich and varied emotions**, such as surprise, excitement, exclamation, and anticipation. To achieve this goal, the transformed dialogue should **include rich colloquial elements** such as interjections, pauses, and modal particles. This text to image prompt is broken down into sentences spoken by human, and make sure that every sentence spoken by the human **does not exceed 15 words**. To achieve this goal, you may first need to discard modifying words, and secondly, you may need to discard some non-critical elements, but please keep the content related to the sentence style, such as modal particles. Make sure that the decomposed sentences spoken by human is coherent, that is, the latter sentence spoken by human is like a supplement to the previous sentences spoken by human.
Require: {ROLE}, {PROMPT}, {MODEL}, {Example 1,...,n}
Input: You are playing the {ROLE}. Please generate natural dialogue based on the following professional {PROMPT} for the {MODEL}. You can refer to following examples: {Example 1,...,n}.

Sample: Single-Input

```
{
"prompt": "1 woman sitting couch closeup modern
↪  living room Andrew Wyeth masterpiece  best
↪  quality dramatic lighting, best shadow",
"freestyle query": "Create a high-quality image of a
↪  woman sitting on a couch in a modern living room,
↪  inspired by Andrew Wyeth's style. Focus on a
↪  close-up view with dramatic lighting and strong
↪  shadows."
}
```

Sample: Multimodal-Input

```
{
"prompt": "a scan of an ID , Altichiero, detailed
↪  product photo, a character portrait a woman in a
↪  shiny suit with pink hair, new objectivity ,
↪  cyberpunk style",
"image" : "<IMAGE_PATH>"
"freestyle query": "Shall we maintain the same level
↪  of detail as the reference images for cyberpunk
↪  aesthetics?"
}
```

Sample: History-Input

```
{
"prompt": "masterpiece, best quality, highres,
↪  gracemiku, garden, stage, smile, standing, cowboy
↪  shot",
"freestyle query":
{
    "round1":
    {
    "freestyle": "Hey, I've got this amazing idea!
    ↪  Picture gracemiku in a beautiful garden
    ↪  stage!",
    "prompt": "masterpiece, best quality, highres,
    ↪  1gracemiku, garden, stage"
    },
    "round2":
    {
    "freestyle": "She's smiling radiantly, and we can
    ↪  see her from the waist up. So gorgeous!",
    "prompt": "masterpiece, best quality, highres,
    ↪  1gracemiku, garden, stage, smile, standing,
    ↪  cowboy shot, radiant, gorgeous"
    },
    "round3":
    {
    "freestyle": "Oh, and make sure to include lots
    ↪  of vivid details. It'll be incredible!",
    "prompt": "masterpiece, best quality, highres,
    ↪  1gracemiku, garden, stage, smile, standing,
    ↪  cowboy shot, radiant, gorgeous, more
    ↪  detailed"
    }
}
}
```

References

[1] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, and Yunxin Jiao. Improving image generation with better captions. 2023. 4

[2] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021. 6

[3] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[4] SpaCy. spacy: Industrial-strength natural language processing in python. `https://spacy.io/`, 2015. 4

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision, 2015. 6

[6] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 6

[7] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023. 6

[8] Lirui Zhao, Yue Yang, Kaipeng Zhang, Wenqi Shao, Yuxin Zhang, Yu Qiao, Ping Luo, and Rongrong Ji. Diffagent: Fast and accurate text-to-image api selection with large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6390–6399, 2024. 6