Hunyuan-DiT : A Powerful Multi-Resolution Diffusion Transformer with Fine-Grained Chinese Understanding

Zhimin Li*, Jianwei Zhang*, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang,

Xingchao Liu, Minbin Huang, Zedong Xiao, Dayou Chen, Jiajun He, Jiahao Li, Wenyue Li, Chen Zhang,

Rongwei Quan, Jianxiang Lu, Jiabin Huang, Xiaoyan Yuan, Xiaoxiao Zheng, Yixuan Li, Jihong Zhang,

Chao Zhang, Meng Chen, Jie Liu, Zheng Fang, Weiyan Wang, Jinbao Xue, Yangyu Tao, Jianchen Zhu,

Kai Liu, Sihuan Lin, Yifu Sun, Yun Li, Dongdong Wang, Zhichao Hu, Xiao Xiao, Yan Chen, Yuhong Liu,

Wei Liu, Di Wang, Yong Yang, Jie Jiang, Qinglin Lu

Tencent Hunyuan

Abstract

We present Hunyuan-DiT , a text-to-image diffusion transformer with fine-grained understanding of both English and Chinese. To construct Hunyuan-DiT , we carefully designed the transformer structure, text encoder, and positional encoding. We also build from scratch a whole data pipeline to update and evaluate data for iterative model optimization. For fine-grained language understanding, we train a Multimodal Large Language Model to refine the captions of the images. Finally, Hunyuan-DiT can perform multi-turn multimodal dialogue with users, generating and refining images according to the context. Through our carefully designed holistic human evaluation protocol with more than 50 professional human evaluators, Hunyuan-DiT sets a new state-of-the-art in Chinese-to-image generation compared with other open-source models.

1 Introduction

Diffusion-based text-to-image generative models, such as DALL-E [5], Stable Diffusion [23, 9] and Pixart [7], have shown the ability to generate images with unprecedented quality. However, they lack the ability to directly understand Chinese prompts, limiting their potential in image generation with Chinese text prompts. To improve Chinese understanding, AltDiffusion [37], PAI-Diffusion [32] and Taiyi [34] were proposed but their generation quality still needs improvement.

In this report, we introduce our entire pipeline for constructing Hunyuan-DiT , which can generate detailed high-quality images in multiple different resolutions according to both English and Chinese prompts. Hunyuan-DiT is made possible by our following efforts: (1) we design a new network architecture based on diffusion transformer [22]. It combines two text encoders, a bilingual CLIP [24] and a multilingual T5 encoder [25] to improve language understanding and increase the context length. (2) we build from scratch a data processing pipeline to add data, filter data, maintain data, update data and apply data to optimize our text-to-image model. Specifically, an iterative procedure called 'data convoy' is designed to examine the effectiveness of new data. (3) we refine the raw captions in the image-text data pairs with Multimodal Large Language Model (MLLM). Our MLLM is fine-tuned to generate structural captions with world

*equal contribution

Figure 1: Hunyuan-DiT can generate images containing Chinese elements.

knowledge. (4) we enable Hunyuan-DiT to interactively modify its generation by having multi-turn dialogues with the user. (5) we perform post-training optimization in the inference stage to lower the deployment cost of Hunyuan-DiT .

To thoroughly evaluate the performance of Hunyuan-DiT , we also created an evaluation protocol with ≥ 50 professional evaluators. The protocol carefully takes into account the different dimensions of a text-to-image model, including text-image consistency, AI artifacts, subject clarity, aesthetics, etc. Our evaluation protocol is incorporated into the data convoy to update the generative model.

Our model, Hunyuan-DiT , achieves state-of-the-art performance among open-source models. In Chinese-to-image generation, Hunyuan-DiT is the best in text-image consistency, excluding AI artifacts, subject clarity, and aesthetics compared with existing open-source models, including Stable Diffusion 3. It performs similarly as top closed-source models, such as DALL-E 3 and MidJourney v6, in subject clarity and aesthetics. Qualitatively, for Chinese elements understanding, including categories such as ancient Chinese poetry and Chinese cuisine, Hunyuan-DiT can generate results with higher image quality and semantic accuracy compared to other comparison algorithms. Hunyuan-DiT supports long text understanding up to 256 tokens.

湖水清澈，天空湛蓝，阳光灿烂。一只优雅的白天鹅在湖边游泳。它周围有几只小鸭子，看起来非常可爱，整个画面给人一种宁静祥和的感觉。
The lake is clear, the sky is blue, and the sun is shining. An elegant white swan swims by the lake. There are several ducklings around it, which looks very cute. The whole picture gives people a feeling of peace and tranquility.

太阳微微升起，花园里的玫瑰花瓣上露珠晶莹剔透，一只瓢虫正在爬向露珠，背景是清晨的花园，微距镜头。
The sun rises slightly, the dew on the rose petals in the garden is crystal clear, a ladybug is crawling to the dew, the background is the early morning garden, macro lens.

一只长靴猫手持亮银色的宝剑，身着铠甲，眼神坚毅，站在一堆金币上，背景是暗色调的洞穴，图像上有金币的光影点缀。
A Puss in Boots, armed with a bright silver sword and clad in armor, stood on top of a pile of gold coins. The background was a dark-colored cave, and the image was dotted with gold coins.

一颗新鲜的草莓特写，光泽鲜艳，颜色诱人，被放在绿色的叶子上，背景是棕色的桌面。
A close-up of a fresh strawberry, glorious, attempting color, lying on pale green leaves with brown table in the background.

一只优雅的白鹤在湖边静静地站立，它的身体纯白色，翅膀轻轻展开，背景是湖面和远处的山脉。
An elegant white crane stands quietly beside the lake. Its body is pure white and its wings are gently spread. The background is the lake and the mountains in the distance.

画面主要描述一个巴洛克风格的花瓶，带有金色的装饰边框，花瓶上盛开着各种色彩鲜艳的花，白色背景。
The picture mainly depicts a Baroque vase with a gold decorative frame, the vase is blooming with various colorful flowers, white background.

一个亚洲女生，头发是黑色，衣服是纯白色短袖，人物风格清新，城市背景。
An Asian girl, black hair, white T-shirt, fresh style, urban background.

一只聪明的狐狸走在阔叶树林里，旁边是一条小溪，细节真实，摄影。
A clever fox walks in a broadleaf forest, next to a stream, detail true, photography.

Figure 2: Hunyuan-DiT can generate images according to fine-grained text prompts.

一幅细致的油画描绘了一只年轻獾轻轻嗅着一朵明亮的黄色玫瑰时错综复杂的皮毛。背景是一棵大树干的粗糙纹理，獾的爪子轻轻地挖进树皮。在柔和的背景中，一个宁静的瀑布倾泻而下，它的水在绿色植物中闪烁着蓝色。
A detailed oil painting captures the intricate fur of a young badger as it gently sniffs at a bright yellow rose. The scene is set against the rough texture of a large tree trunk, with the badger's claws slightly digging into the bark. In the softly painted background, a tranquil waterfall cascades down, its waters a shimmering blue amidst the greenery.

一个亚洲中年男士在夕阳下的公园长椅上静坐。他穿着一件深蓝色的针织毛衣和灰色裤子。他的头发略显花白，手中拿着一本敞开的书。面带微笑，眼神温和，周围是落日余晖和四周的绿树。
A middle-aged Asian man sits on a park bench at sunset. He was wearing a dark blue knit sweater and grey trousers. His hair was slightly graying and he held an open book in his hand. He smiled, his eyes were gentle, and he was surrounded by sunset and green trees.

一张细致的照片捕捉到了一尊雕像的形象，这尊雕像酷似一位古代法老，头上出人意料地戴着一副青铜蒸汽朋克护目镜。这座雕像穿着复古时髦，清爽的白色T恤和合身的黑色皮夹克，与传统的头饰形成鲜明对比。背景是简单的纯色，突出了雕像的非传统服装和蒸汽朋克眼镜的复杂细节。
A detailed photograph captures the image of a statue with the likeness of an ancient pharaoh, unexpectedly accessorized with a pair of bronze steampunk goggles resting atop its head. The statue is dressed in an anachronistic fashion, featuring a crisp white t-shirt and a fitted black leather jacket that contrasts with its traditional headdress. The background is a simple, solid color that accentuates the statue's unconventional attire and the intricate details of the steampunk eyewear.

一位年轻女子站在春季的火车站月台上。她身着蓝灰色长风衣，白色衬衫。她的深棕色头发扎成低马尾，几缕碎发随风飘扬。她的眼神充满期待，阳光洒在她温暖的脸庞上。
A young woman stands on the platform of a railway station in spring. She wore a long blue-gray windbreaker and white shirt. Her dark brown hair was tied in a low ponytail, a few strands flying in the wind. Her eyes were full of expectation, and the sun shone on her warm face.

一个异想天开的场景，一只美洲驼，戴着一副超大的圆形太阳镜，自信地站在宇宙飞船的金属甲板上。美洲驼蹄子下的甲板闪闪发光，抛光的银色，反映了围绕着船只的星空。在广阔的背景下，地球若隐若现，蓝色的海洋和白色的云层形成了漩涡，与飞船的时尚，未来主义的设计形成了惊人的对比。
A whimsical scene where a llama, adorned with a pair of oversized, round sunglasses, stands confidently on the metallic deck of a spacecraft. The deck beneath the llama's hooves gleams with a polished silver finish, reflecting the starry cosmos that surrounds the vessel. In the vast backdrop, the Earth looms large, a swirl of blue oceans and white clouds, providing a stunning contrast to the spaceship's sleek, futuristic design.

Figure 3: Hunyuan-DiT can generate images following long text prompts.

鲤鱼跃龙门。
The carp leaps over dragon gate.

臭豆腐。
Stinky Tofu.

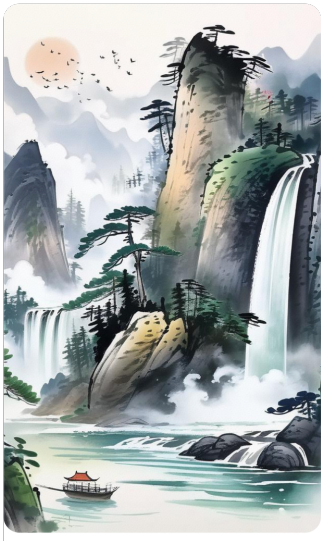
乌镇风光。
The view of Wuzhen.

一只可爱的猫, 细节真实, 摄影。
A cute cat, true details, photography.

一朵鲜艳的红色玫瑰花，花瓣撒有一些水珠，晶莹剔透，特写镜头。
A bright red rose, petals scattered with some water droplets, crystal clear, close-up.

泥塑风格，一座五彩斑斓的花园在画面中展现，各种各样的花朵，绿色的叶子和一只正在嬉戏的小猫形成了一幅生动的图像，背景是蓝天和白云。
Clay sculpture style, a colorful garden is displayed in the picture, various flowers, green leaves and a cat playing form a vivid image, the background is blue sky and white clouds.

夏日黄昏，在薰衣草田的一角，一条用旧木板搭建的小路通向远处的一座小木屋，而在小路的入口处，一辆自行车被随意地停放着，自行车的篮子里放着刚采摘的薰衣草。
Summer dusk, in a corner of a lavender field, a path built with old wooden boards leads to a small wooden house in the distance. At the entrance of the path, a bicycle is parked casually, with freshly picked lavender in the basket.



飞流直下三千尺，疑是银河落九天。
The waterfall flying down 3,000 feet, seems to be the Milky Way falling from the sky.

Figure 4: Hunyuan-DiT can generate images in various resolutions.

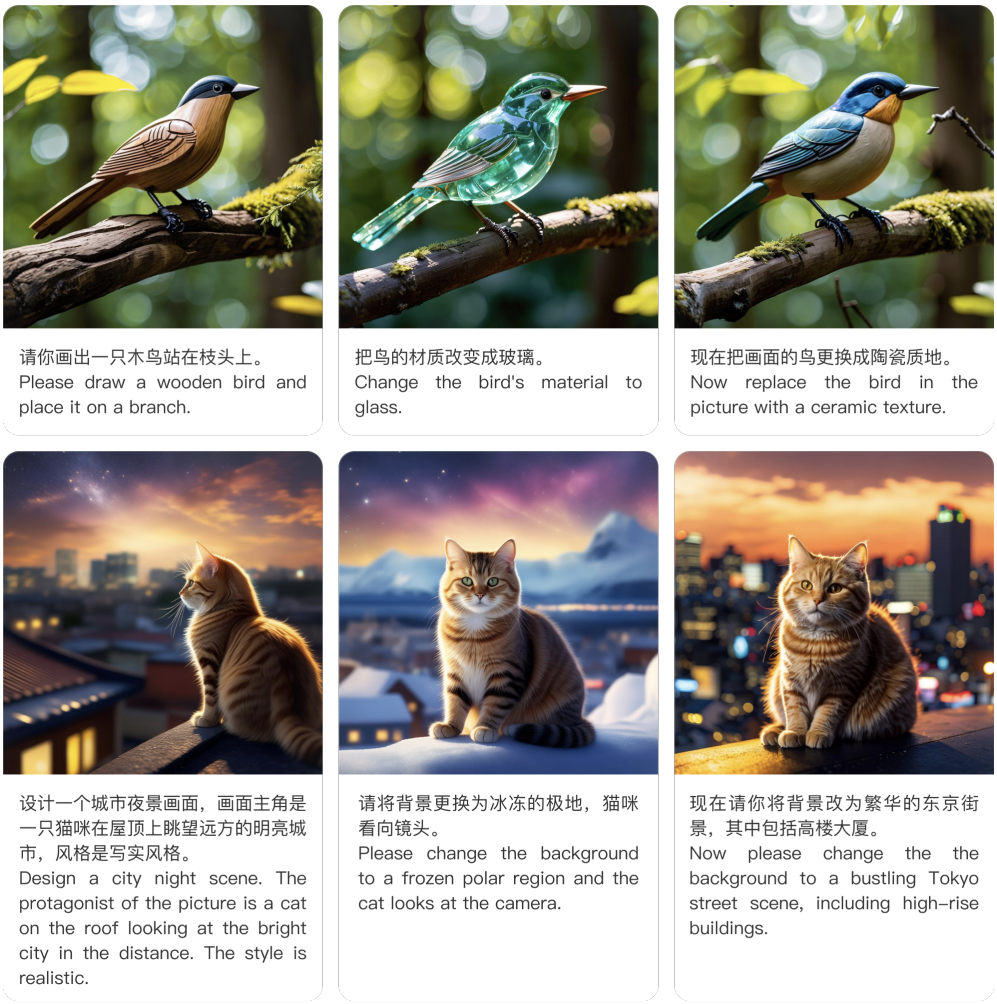


Figure 5: Hunyuan-DiT can generate images in multi-turn dialogue.

2 Methods

2.1 Improved Generation with Diffusion Transformers

Hunyuan-DiT is a diffusion model in the latent space, as depicted in Figure 7. Following the Latent Diffusion Model [26], we use a pre-trained Variational Autoencoder (VAE) to compress the images into low-dimensional latent spaces and train a diffusion model to learn the data distribution with diffusion models. Our diffusion model is parameterized with a transformer [31, 22, 4]. To encode the text prompts, we leverage a combination of pre-trained bilingual (English and Chinese) CLIP [24] and multilingual T5 encoder [25]. We will introduce the details of each module in sequel.

VAE We use the VAE in SDXL [23], which is fine-tuned on 512 $\times$ 512 images from the VAE in SD 1.5 [26]. Experimental findings show that the text-to-image models trained on the high-resolution SDXL VAE improved clarity, alleviated over-saturation, and reduced distortions over SD 1.5 VAE. As the VAE latent space greatly influences generation quality, we will explore a better training paradigm for the VAE in the future.

The Diffusion Transformer in Hunyuan-DiT Our diffusion transformer has several improvements compared to the baseline DiT [22]. We found the Adaptive Layer Norm used in class-conditional DiT performs unsatisfactorily to enforce fine-grained text conditions. Therefore, we modify the model structure to combine the text condition with the diffusion model using cross-attention as Stable Diffusion [26]. Hunyuan-DiT takes a vector $x \in \mathbb{R}^{c\times h\times w}$ in the latent space of the VAE as input, and then patchifies x into $\frac{h}{p} \times \frac{w}{p}$ patches, where p is set to 2. After a linear projection layer, we get $hw/4$ tokens for the subsequent transformer blocks. Hunyuan-DiT has two types of transformer

blocks, the encoder block and the decoder block. Both of them contain three modules - self-attention, cross-attention, and feed-forward network (FFN). The text information is fused in the cross-attention module. The decoder block additionally contains a skip module, which adds the information from the encoder block in the decoding stage. The skip module is similar to the long skip-connection in U-Nets, but there are no upsampling or downsampling modules in Hunyuan-DiT due to our transformer structure. Finally, the tokens are reorganized to recover the two-dimensional spatial structure. For training, we find using v-prediction [28] gives better empirical performance.

Text Encoder An efficient text encoder is crucial in text-to-image generation, as they need to accurately understand and encode the input text prompts to generate corresponding images. CLIP [24] and T5 [25] have become the mainstream choices for these encoders. Matryoshka diffusion models [11], Imagen [27], MUSE [6], and Pixart-α [7] use solely T5 to enhance their understanding of the input text prompts. In contrast, eDiff-I [3] and Swinv2-Imagen [17] fuse the two encoders, CLIP and T5, to further improve their text understanding capabilities. Hunyuan-DiT chooses to combine T5 and CLIP in text encoding to leverage the advantages of both models, thereby enhancing the accuracy and diversity of the text-to-image generation process.

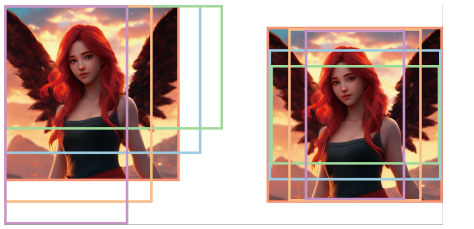


Figure 6: Illustration of Extended Positional Encoding and Centralized Interpolative Positional Encoding

Positional Encoding and Multi-Resolution Generation A common practice in visual transformers [22, 8] is to apply sinusoidal positional encoding that encodes the absolute position of a token. In Hunyuan-DiT , we employ the Rotary Positional Embedding (RoPE) [30] to simultaneously encode the absolute position and relative position dependency. We use two-dimensional RoPE which extends RoPE to the image domain.

Hunyuan-DiT supports multi-resolution training and inference, which requires us to assign appropriate positional encodings for different resolutions. For $x \in \mathbb{R}^{c\times h\times w}$, we tried two types of positional encoding for multi-resolution generation:

1. **Extended Positional Encoding:** Extended Positional Encoding gives the positional encoding of x in a naive way, which is,
$$\text{PE}(x_{i,j}) = (f(i), f(j)), \quad i \in \{1, \cdots, h\}, j \in \{1, \cdots, w\}, \tag{1}$$
where f is the positional encoding function for each coordinate i and j. $\text{PE}(x)$ is the obtained 2D positional encoding for the position (i, j). Note that when the data x has different resolutions, their h and w exhibit huge differences and the positional encoding varies significantly.
2. **Centralized Interpolative Positional Encoding:** We use Centralized Interpolative Positional Encoding to align the positional encoding for x with different h and w. Assuming $h \geq w$, Centralized Interpolative Positional Encoding computes the positional encoding as,
$$\text{PE}(x_{i,j}) = \left(f\left(\frac{S}{2} + \frac{S}{h}\left(i - \frac{h}{2}\right)\right), f\left(\frac{S}{2} + \frac{S}{h}\left(j - \frac{w}{2}\right)\right)\right), \tag{2}$$
where $i \in \{1, \cdots, h\}, j \in \{1, \cdots, w\}$ and S is a pre-defined boundary of the positional encoding. This strategy ensures images with various resolutions to have the same range $[0, S]$ when computing positional encoding, therefore improving the efficiency of learning.

Although Extended Positional Encoding is easier to implement, we observe that it is a suboptimal choice for multi-resolution training. It could not align images with different resolutions or cover the rare cases where both h and w are large. On the contrary, Centralized Interpolative Positional Encoding allows images with different resolutions to share similar positional encoding spaces. With Centralized Interpolative Positional Encoding, the model converges faster and generalizes to new resolutions.

Improving Training Stability To stabilize training, we present three techniques:

1. We add layer normalization in all the attention modules before computing Q, K, and V. This technique is called QK-Norm, which is proposed in [12]. We found it effective for training Hunyuan-DiT as well.
2. We add layer normalization after the skip module in the decoder blocks to avoid loss explosion during training.
3. We found certain operations, e.g., layer normalization, tend to overflow with FP16. We specifically switch them to FP32 to avoid numerical errors.

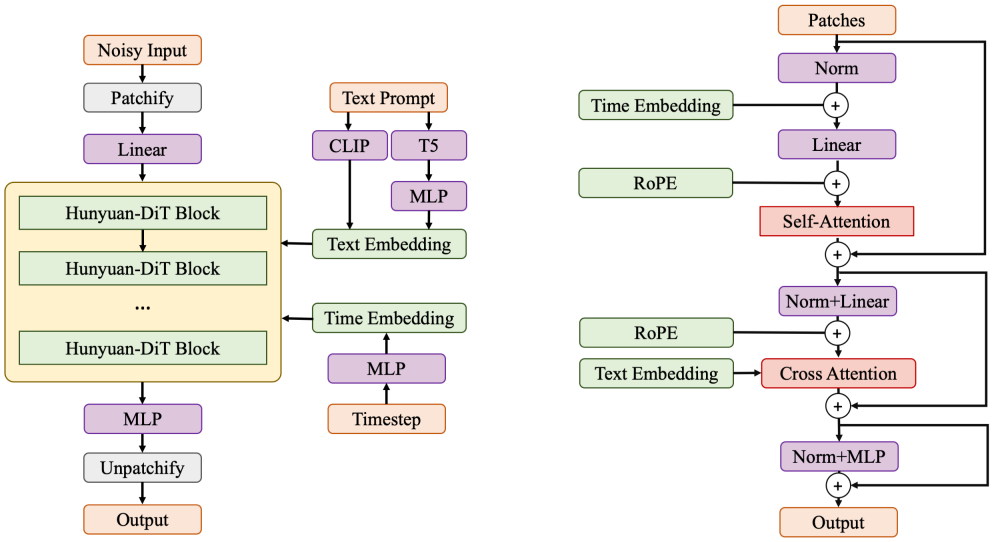


Figure 7: The model structure of Hunyuan-DiT .

2.2 Data Pipeline

Data Processing The pipeline to preparing our training data is composed of four parts, which is illustrated in Fig.20

1. **Data Acquisition:** The primary channels for data acquisition are currently external purchasing, open data downloading, and authorized partner data.
2. **Data Interpretation:** After obtaining the raw data, we tag the data to identify the strengths and weaknesses of the data. Currently, over ten tagging capabilities are supported, including image clarity, aesthetics, indecency, violence, sexual content, presence of watermarks, image classification, and image description.
3. **Data Layering**: Data layering is constructed for large quantities of images to serve the different stages of model training. For example, billions of image-text pairs are used as copper-tier data to train our foundational CLIP model [24]. Then, a relatively high-quality image set is screened from this large library as silver-tier data to train the generative model to improve the model's quality and understanding capabilities. Lastly, through machine screening and manual annotation, the highest quality data is selected as gold-tier data for refining and optimizing the generative model.
4. **Data Application**: The hierarchical data are applied to several areas. Specialized data is filtered out for specialty optimizations, e.g, person or style specializations. Newly processed data is continually added to the iterative optimization of the foundation generative model. The data is also frequently inspected to maintain the quality of the ongoing data processing.

Data Category System We found the coverage of the data categories in the training data crucial for training accurate text-to-image models. Here we discuss two fundamental categories:

1. **Subject:** The generation of the subject is the foundational ability of the text-to-image model. Our training data covers a vast majority of categories, including human, landscape, plants, animals, goods, transportation, games, and more, with over ten thousand sub-categories.
2. **Style:** The diversity of the style is critical to the user's preference and stickiness. Currently, we have covered over a hundred styles, including anime, 3D, painting, realistic, and traditional styles.

Data Evaluation To evaluate the impact of introducing specialized data or newly processed data on the generative model, we design a 'data convoy' mechanism, depicted in Fig.21, which is composed of:

1. We categorize the training data according to the data category system, containing subject, style, scene, composition, etc. Then we adjust the distribution between different categories to meet the model's demand and fine-tune the model with the category-balanced dataset.

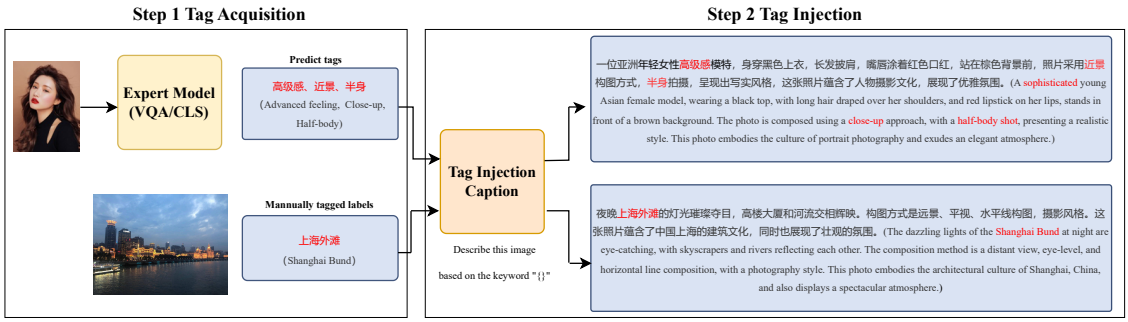


Figure 8: Re-captioning with tag injection based on manual labeling and expert models.

2. We perform category-level comparisons between the fine-tuned model and the original model to evaluate the advantages and drawbacks of the data, relying on which we set the directions to update our data.

Successfully running the mechanism requires a complete evaluation protocol on the text-to-image model. Our model evaluation protocol is composed of two parts:

1. **Evaluation Set Construction:** We construct the initial evaluation set by combining bad cases and business needs based on our data categories. Through human annotation of the reasonableness, logic, and comprehensiveness of the test cases, the usability of the evaluation set is assured.
2. **Evaluation in Data Convoy:** In every data convoy, we randomly select a subset of test cases from the evaluation set to form a holistic evaluation subset including subjects, styles, scenes, compositions. We compute an overall score of all the evaluated dimensions to assist the iteration of data.

We will elaborate our evaluation protocol in Section 3.

2.3 Caption Refinement for Fine-Grained Chinese Understanding

The image-text pairs obtained from crawling the Internet are usually low-quality pairs, and improving the corresponding captions for the images is important for training text-to-image models [7, 5]. Hunyuan-DiT adopts a well-trained multimodal large language model (MLLM) to re-caption the raw image-text pairs to enhance the data quality. We adopt structural captions to comprehensively describe the images. Furthermore, we also use raw captions and expert models that include world knowledge to enable the generation of special concepts in the re-captioning.

Re-captioning with Structural Captions Existing MLLMs, e.g., BLIP-2 [16] and Qwen-VL [2] tend to generate over-simplified captions that resemble MS-COCO captions [18] or highly redundant captions that are not related to the images. To train an MLLM that is suitable to improve raw image-text pairs, we construct a large-scale dataset for structural captions and fine-tune the MLLM.

We use an AI-assisted pipeline for dataset construction. Human labeling for image captioning is difficult, and the labeling quality can hardly be standardized. Therefore, we use a three-stage pipeline to boost labeling efficiency with AI assistance. In Stage 1, we ensemble the captions from multiple basic image captioning models with human labeling to get an initial dataset. In Stage 2, we train the MLLM with the initial dataset, and then use the trained model to generate new captions for the images. As its re-captioning accuracy is enhanced, the efficiency of human labeling is improved by around 4 times.

Our model structure is similar to LLAVA-1.6 [19]. It is composed of a ViT for vision, a decoder-only LLM for language, and an Adapter for bridging vision and text. The training objective is the classification loss as other auto-regressive models.

Re-captioning with Information Injection In human labeling of structural captions, world knowledge is always missing because it is impossible for human to recognize all the special concepts in the images.

We leverage two methods to inject world knowledge to the captions:

1. **Re-captioning with Tag Injection:** To simplify the labeling process, we can label tags of images and use MLLMs to generate tag-injected captions from the labeled tags. Besides labeling with human experts, we can use expert models to get the tags, including but not limited to general object detectors, landmark classification

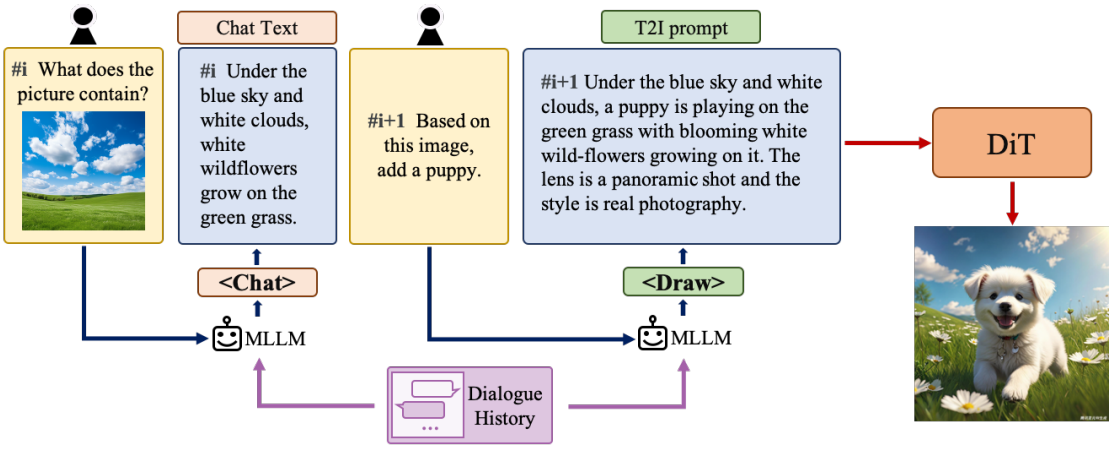


Figure 9: Our pipeline of text-to-image generation with multi-turn dialogue.

models, and action recognition models. The additional information from tags can significantly add to the world knowledge in the generated captions. To this end, we design an MLLM that takes images and tags as input and outputs more comprehensive captions containing the information from the tags. We found this MLLM can be trained with very sparse human-labeled data.

2. **Re-captioning with Raw Captions:** Capsfusion [39] proposed to fuse raw captions with generated descriptive captions using ChatGPT. However, raw captions are usually noisy and LLM alone cannot correct the wrong information in the raw captions. To alleviate this, we construct an MLLM that generates captions from both images and raw captions, which can correct the mistakes by taking image information into account.

2.4 Prompt Enhancement with Multi-Turn Dialogue

Understanding natural language instructions and performing multi-turn interactions with users are important for a text-to-image system. It can help build a dynamic and iterative creation process that brings the user's idea into reality step by step. In this section, we will detail how we empower Hunyuan-DiT with the ability to perform multi-turn conversations and image generation. Various works have made efforts to equip text-to-image models with the multi-turn ability using MLLMs, such as Next-GPT [33], SEED-LLaMA [10], RPG [36], and DALLE-3 [5]. These models either use the MLLM to generate text prompts or the text embeddings for the text-to-image model. We choose the first choice as generating text prompts is more flexible. We train MLLM to understand the multi-turn user dialogue and output the new text prompt for image generation.

Text Prompt Enhancement Natural language instructions given by the user have a huge difference with the refined captions on which the text-to-image generative model is trained. Consequently, we need a model to transform these instructions into detailed semantically coherent text prompts for successful high-quality image generation. To train this model, we use the in-context learning ability of GPT-4. We collect a small set of manually annotated `(instruction, text prompt)` pairs as in-context learning examples, then we query GPT-4 to generate more data pairs. These pairs construct a single-turn instruction-to-prompt dataset, referred to as D_p.

Multimodal Multi-Turn Dialogue Normal MLLMs only support text output. To align with our goal to build a multi-turn text-to-image generation system, we add a special token `<draw>` to indicate that a text prompt should be sent to Hunyuan-DiT in the current turn of conversation. If the model successfully predicts the `<draw>` token, it will generate a detailed prompt for Hunyuan-DiT . To train the MLLM, we design a dataset of three-turn multimodal conversations. To ensure broad coverage of conversational scenarios, we explore different combinations of input and output types based on four primary categories, i.e., text $\rightarrow$ text, text $\rightarrow$ image, text+image $\rightarrow$ text, text+image $\rightarrow$ image. By selecting a type in each turn of conversation, we pre-define a set of three-turn dialogue compositions. For each composition, we then employ GPT-4 to generate the 'dialogue prompts', which are used to define the behavior of the AI agent before the dialogue, leading to unique conversational flows. We traverse 13 topics and 7 image editing methods to yield $\sim$15,000 samples after querying GPT-4 with various 'dialogue prompts'. In the 'dialogue prompts', we also add the samples in D_p to avoid the distribution shift of the generated text prompts. We denote this dataset of three-turn text-to-image conversations as D_{tt}.

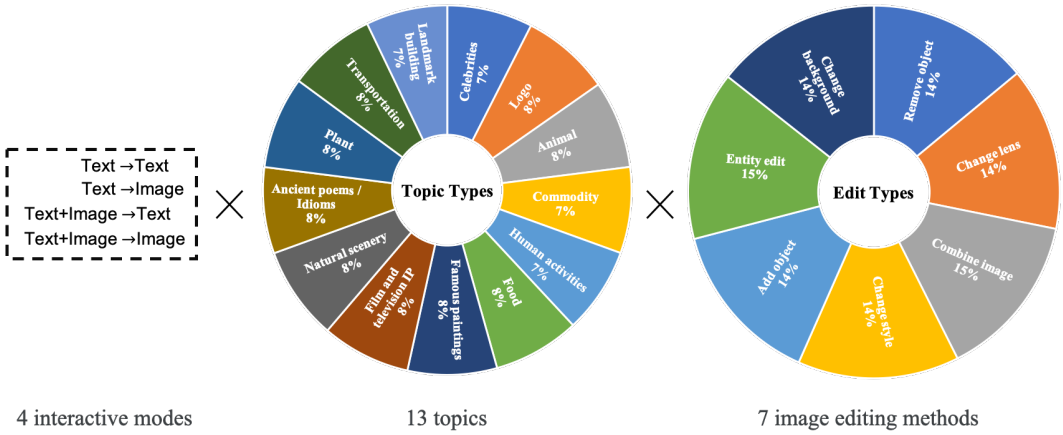


Figure 10: Data construction for multi-turn dialogue.

Instruction Tuning Data Mixing To maintain the multimodal conversation ability, we also included a range of open-sourced uni/multimodal conversation datasets, denoted as D_o. We randomly shuffle and concatenate the single-turn samples from D_p and D_o to get a pseudo-multi-turn dataset D_{pm}. This dataset features multi-turn conversations but not necessarily preserving semantic coherence, simulating the scenarios in which the user may switch the topic within a conversation. To accommodate change of topic, we train the model to predict a `<switch>` token. We mix the collection of D_o, D_p, D_{pm} together with D_{tt} to serve as the final training dataset D. For more details, please refer to [14].

Guarantee on Subject Consistency In multi-turn text-to-image, users may ask the AI system to edit a certain subject multiple times. Our goal is to ensure that the subjects generated across multiple conversational turns remain as consistent as possible. To achieve this, we add the following constraints in the 'dialogue prompts' of the dialogue AI agent. For image generation that builds upon the images produced in previous turns, the transformed text prompts should satisfy the user's current demand while being altered as little as possible from the text prompts used for previous images. Moreover, during the inference phase of a given conversation, we fix the random seed of the text-to-image model. This approach significantly increases the subject consistency throughout the dialogue.

2.5 Optimization in the Inference Stage

Engineering Optimization Deploying Hunyuan-DiT for the users is expensive, we adopt multiple engineering optimization strategies to improve the inference efficiency, including ONNX graph optimization, kernel optimization, operator fusion, precomputation, and GPU memory reuse.

Algorithmic Acceleration Recently, various methods have been proposed to reduce the inference steps of diffusion-based text-to-image models [21, 29, 20, 35, 38]. We attempted to apply these methods to accelerate Hunyuan-DiT , and the following problems arise:

1. **Training Stability:** We observed adversarial training tends to collapse due to the unstable training scheme.
2. **Adaptivity:** We found several methods results in models that cannot reuse the pre-trained plug-in modules or LoRAs.
3. **Flexibility:** In our practice, the Latent Consistency Model is only suitable for low-step generation. Its performance deteriorates when the number of inference steps increases beyond a certain threshold. This limitation prevents us from flexibly adjusting the balance between generation performance and acceleration.
4. **Training Cost:** Adversarial training introduces additional modules for training the discriminative model, which brings severe demand of extra GPU memory and training time.

Considering these problems, we choose Progressive Distillation [28]. It enjoys stable training and allows us to smoothly trade-off between the acceleration ratio and the performance, offering us the cheapest and fastest way for model acceleration. To encourage the student model to accurately imitate the teacher model, we carefully tune the optimizer, classifier-free guidance, and regularization in the training process.

3 Evaluation Protocol

To holistically evaluate the generation ability of Hunyuan-DiT , we constructed a multi-dimensional evaluation protocol, which is composed of evaluation metrics, evaluation dataset construction, evaluation execution, and evaluation protocol evolution.

3.1 Evaluation Metrics

Evaluation Dimensions When determining the evaluation dimensions, we referenced existing literature and additionally invited professional designers and general users to participate in interviews to ensure that the evaluation metrics have both professionalism and practicality. Specifically, when evaluating the capabilities of our text-to-image models, we adopted the following four dimensions: text-image consistency, AI artifacts, subject clarity, and overall aesthetics. For results that raise safety concerns (such as involving pornography, politics, violence, or bloodshed), we directly mark them as `unacceptable`.

Multi-Turn Interaction Evaluation When evaluating the capabilities of the multi-turn dialogue interaction, we also assessed extra dimensions such as instruction compliance, subject consistency, and the performance of multi-turn prompt enhancement for image generation.

3.2 Evaluation Dataset Construction

Dataset Construction We combine AI-generated and human-created test prompts to construct a hierarchical evaluation dataset with various difficulty levels. Specifically, we categorize the evaluation dataset into three difficulty levels - easy, medium, and hard - based on factors such as the richness of the text prompt content, the number of descriptive elements (main subject, subject modifiers, background descriptions, styles, etc.), whether the elements are common, and whether they contain abstract semantics (e.g. poems, idioms, proverbs).

Furthermore, due to the issues of homogeneity and long production cycles when creating test prompts with humans, we rely on LLMs to enhance the diversity and difficulty of the test prompts, rapidly iterate on prompt generation, and reduce manual labor.

Evaluation Dataset Categories and Distribution In the process of constructing hierarchical evaluation dataset, we analyzed the text prompts used by users when using the text-to-image generative models, and combined user interviews and expert designer opinions to cover functional applications, character roles, Chinese elements, multi-turn text-to-image generation, artistic styles, subject details, and other major categories in the evaluation dataset.

The different categories are further divided into multiple hierarchical levels. For example, the 'subject details' category is further divided into subcategories like animals, plants, vehicles, and landmarks. For each subcategory, we maintain a prompt count of more than 30.

3.3 Evaluation Execution

Evaluation Team The evaluation team consists of professional evaluators. They have rich professional knowledge and evaluation experience, allowing them to accurately execute the evaluation tasks and provide in-depth analysis. The evaluation team has more than 50 members.

Evaluation Process The evaluation process includes two stages: evaluation standard training and multi-person correction. In the evaluation standard training stage, we provide detailed training to the evaluators to ensure they have a clear understanding of the evaluation metrics and the tools. In the multi-person correction stage, we have multiple evaluators independently evaluate the same set of images, then summarize and analyze the evaluation results to mitigate subjective biases among the evaluators.

Particularly, the evaluation dataset was structured in a 3-level hierarchical manner, with 8 level-1 categories and more than 70 level-2 categories. For each level-2 category, we have 30 - 50 prompts in the evaluation set. The evaluation set has more than 3,000 prompts in total. Specifically, our evaluation score is computed with the following steps:

1. **Calculating Results for Individual Prompts:** For each prompt, we invite multiple evaluators to independently assess the images generated by the model. We then aggregate the evaluators' assessments and calculate the percentage of evaluators who consider the image to be acceptable. For example, if 10 evaluators are involved and 7 of them consider the image acceptable, the pass rate for that prompt is 70%.

Type	Model	Text-Image Consistency (%)	Excluding AI Artifacts (%)	Subject Clarity (%)	Aesthetics (%)	Overall (%)
Open	**Hunyuan-DiT**	**74.2**	**74.3**	**95.4**	**86.6**	**59.0**
	Playground 2.5 [15]	71.9	70.8	94.9	83.3	54.3
	PixArt-α [7]	68.3	60.9	93.2	77.5	45.5
	SDXL [23]	64.3	60.6	91.1	76.3	42.7
Closed	DALL-E 3 [5]	83.9	80.3	96.5	89.4	71.0
	SD 3 [9]	77.1	69.3	94.6	82.5	56.7
	MidJourney v6 [1]	73.5	80.2	93.5	87.2	63.3

Table 1: Comparison with other state-of-the-art models. **Bold** refers to the highest score in open-source models.

2. **Calculating Level-2 Category Scores:** We classify the prompts into level-2 categories according to their contents. Each prompt within the same level-2 category has equal weight. For all the prompts under the same level-2 category, we calculate the average of their pass rates to obtain the score for that level-2 category. For example, if a level-2 category has 5 prompts with pass rates of 60%, 70%, 80%, 90%, and 100%, the score for that level-2 category is (60% + 70% + 80% + 90% + 100%) / 5 = 80%.
3. **Calculating Level-1 Category Scores:** Based on the level-2 category scores, we calculate the scores for the level-1 categories. For each level-1 category, we take the average of the scores of its subordinate level-2 categories to obtain the level-1 category score. For example, if a level-1 category has 3 level-2 categories with scores of 70%, 80%, and 90%, the level-1 category score is (70% + 80% + 90%) / 3 = 80%.
4. **Calculating the Overall Pass Rate:** Finally, we calculate the overall pass rate based on the weights of each level-1 category. Suppose there are 3 level-1 categories with scores of 70%, 80%, and 90%, and weights of 0.3, 0.5, and 0.2 respectively, the overall pass rate would be $0.3 \times 70\% + 0.5 \times 80\% + 0.2 \times 90\% = 79\%$. The weights of the level-1 categories are determined by careful discussion with users, designers and experts, as shown in Table 2.

Through the above process, we can obtain the pass rates of the model at different category levels, as well as the overall pass rate, to comprehensively evaluate the model's performance.

Evaluation Result Analysis After evaluation, we conduct in-depth analysis of the results, including:

1. Comprehensive analysis of the results for different evaluation metrics (text-image consistency, AI artifacts, subject clarity, and overall aesthetics) to understand the model's performance in various aspects.
2. Comparative analysis of the model's performance on tasks of different difficulty levels to understand the model's capabilities in handling complex scenarios and abstract semantics.
3. Identifying the model's strengths and weaknesses to provide directions for future optimization.
4. Comparison with other state-of-the-art models.

3.4 Evaluation Protocol Evolution

In the continuous optimization of the evaluation framework, we will consider the following aspects: To improve our evaluation protocol to accommodate new challenges, we consider the following aspects: (1) introducing new evaluation dimensions; (2) adding in-depth analysis in the evaluation feedback, such as the spots where the text-image inconsistency occurs, or precise markings of distortion locations; (3) dynamically adjusting the evaluation datasets; (4) improving evaluation efficiency by using machine evaluations.

4 Results

4.1 Quantitative Evaluation

Comparison with State-of-the-Art We compared Hunyuan-DiT with state-of-the-art models, including both open-source models (Playground 2.5, PixArt-α, SDXL) and closed-source models (DALL-E 3, SD 3, MidJourney v6). We follow the evaluation protocol in Section 3. All the models are evaluated on four dimensions, including text-image

一位焦虑的中年男子在深夜的道路上迷路(An anxious middle-aged man gets lost on a road late at night.)

画面主要描述一个中年男子站在深夜的道路上，表情焦虑，他看着手中的地图，但眼神却带着迷茫，背景是漆黑的夜空和道路(The scene mainly depicts a middle-aged man standing on a road late at night, with an anxious expression on his face. He is looking at a map in his hand, but his eyes are confused. The background is the dark night sky and the road.)

画一个夸父追日(Draw a picture of Kuafu chasing the sun.)

中国水墨画风，画面主要描述夸父奋力向前跑，汗水如雨下，远处的太阳金光灿烂，在无边的草原上传来明亮的阳光，画面以近景镜头捕捉(In the style of Chinese ink painting, the picture mainly depicts Kuafu running forward with all his might, his sweat falling like rain. The distant sun is shining brilliantly, casting bright sunlight over the boundless grassland. The scene is captured through a close-up shot.)

Figure 11: The effect of prompt enhancement. When it comes to simple abstract concept prompts, prompt enhancement with MLLM can effectively boost the consistency between generated images and their corresponding text descriptions.

Prompt: A steaming basket of Goubuli buns on a tabletop, a high-definition photograph with a close-up shot(一笼狗不理包子冒着热气放在桌面上，高清摄影图片，特写镜头)

Figure 12: Qualitative comparison between Hunyuan-DiT and other SOTA models.

Prompt: The scene mainly depicts a Kunqu artist performing, dressed in a gorgeous costume and wearing exquisite makeup on her face (画面主要描述一个昆曲艺术家正在表演，她穿着华丽的戏服，脸上画着精致的妆容)

Figure 13: Qualitative comparison between Hunyuan-DiT and other SOTA models.

consistency, the ability of excluding AI artifacts, subject clarity, and aesthetics. As depicted in Table 1, Hunyuan-DiT achieves the best score on all the four dimensions compared with other open-source models. In comparison with closed-source models, Hunyuan-DiT can achieve similar performance to SOTA models such as MidJourney v6 and DALL-E 3 in terms of subject clarity and image aesthetics. In terms of the overall pass rate, Hunyuan-DiT ranks third among all models, better than existing open-source alternatives.

4.2 Ablation Study

Experiment Setting Following the setting in prior research [27, 3], we evaluate different variants of the models using the zero-shot Frechet Inception Distance (FID) [13] on MS COCO [18] 256×256 validation dataset by generating 30,000 images from the prompts in the validation set. We also report the average CLIP [24] score of these generated images to examine the correspondence between text prompts and images.

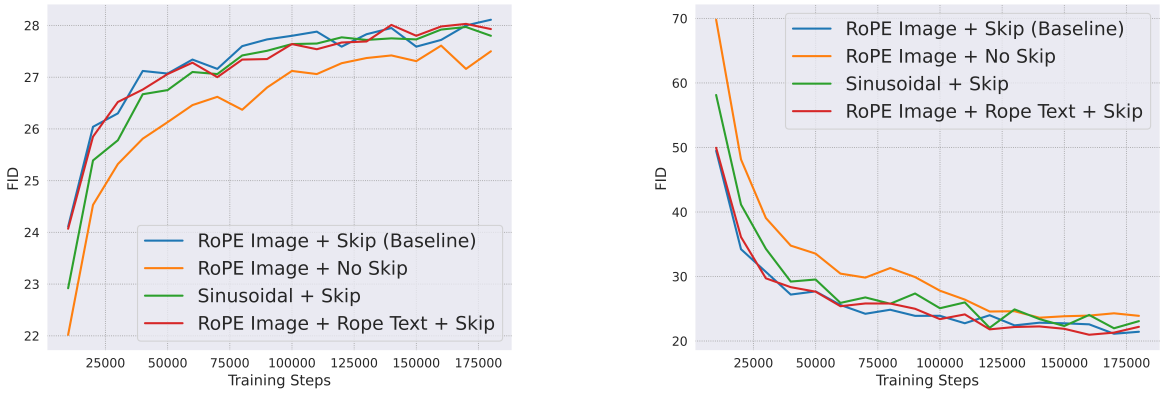
Effect of the Skip Module Long skip connections are utilized to achieve feature fusion between symmetrically positioned encoding and decoding layers in U-Nets. We use Skip Modules in Hunyuan-DiT to mimic this design. As depicted in Figure. 15, we observed that removing long skip connection increases FID and decreases the CLIP score.

Rotary Position Encoding (RoPE) We compare sinusoidal position encoding (the original position encoding in DiT [22]) with RoPE [30]. The results are shown in Figure. 15 as well. We found RoPE position encoding outperformed the sinusoidal position encoding in most time of the training stage. Especially, we found RoPE accelerates the convergence of the model. We hypothesize that this is due to RoPE's ability to encapsulate both absolute and relative positional information.

We also evaluated the inclusion of one-dimensional RoPE position encoding in the text features, as shown in the Figure. 15. We found that adding RoPE position encoding to the text embeddings did not yield significant gains.

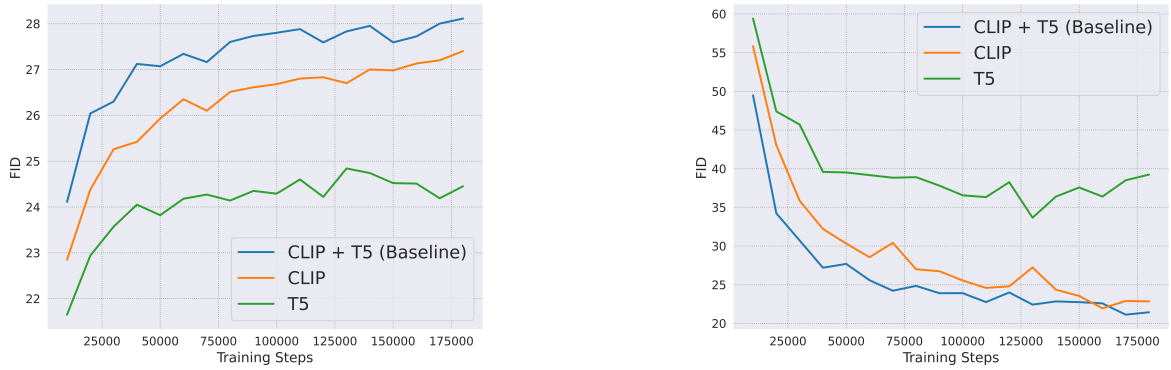
Prompt: A detailed photograph captures the image of a statue resembling an ancient pharaoh, unexpectedly wearing a pair of bronze steampunk goggles. The statue is dressed in outdated fashion, wearing a crisp white T-shirt and a fitted black leather jacket, contrasting with the traditional headdress. The background is a simple solid color, highlighting the intricate details of the statue's unconventional clothing and steampunk eyewear.(一尊雕像的细致的照片形象，这尊雕像酷似一位古代法老，出人意料地戴着一副青铜蒸汽朋克护目镜。这座雕像穿着过时的时尚，穿着一件清爽的白色T恤和一件合身的黑色皮夹克，与传统的头饰形成鲜明对比。背景是简单的纯色，突出了雕像的非传统服装和蒸汽朋克眼镜的复杂细节)

Figure 14: Qualitative comparison between Hunyuan-DiT and other SOTA models.



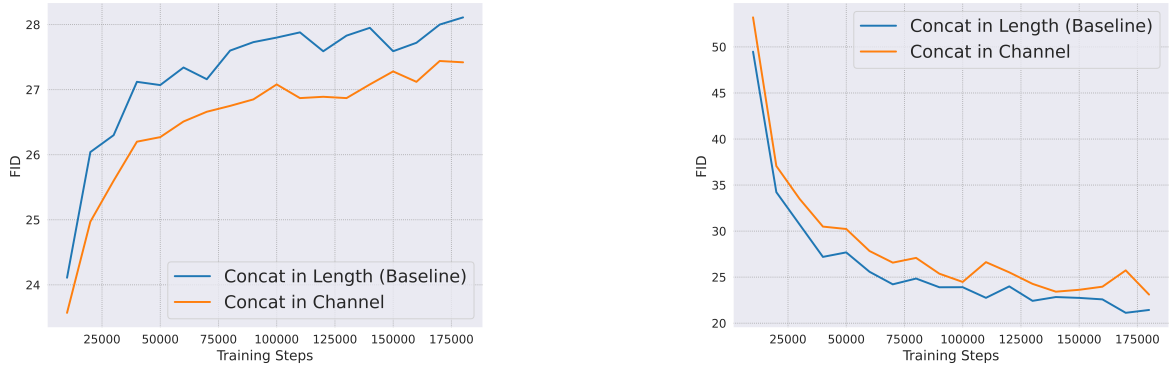
(a) CLIP Score (b) FID

Figure 15: Ablation study on position encoding and model structure.



(a) CLIP Score (b) FID

Figure 16: Ablation study on different schemes of text encoding.



(a) CLIP Score (b) FID

Figure 17: Ablation study on concatenating features of the text encoders.

Text Encoder We evaluated three schemes for text encoding: (1) using our own bilingual (Chinese-English) CLIP alone, (2) using multilingual T5 alone, and (3) using both bilingual CLIP and multilingual T5. In Figure. 16, using CLIP encoder alone outperforms using multilingual T5 encoder alone. Moreover, combining the bilingual CLIP encoder with multilingual T5 encoder leverages both the efficient semantic capture ability of CLIP and the fine-grained semantic understanding advantage of T5, leading to a significantly enhanced FID and CLIP score.

We also explored two ways of concatenating features from CLIP and T5 in Figure. 17: merging along the channel dimension and merging along the length dimension. We found that concatenating the features of text encoders along the text length dimension yields superior performance. Our hypothesis is that, by concatenating along the text length dimension, the model can fully leverage the Transformer's global attention mechanism to focus on each text slot. This facilitates a better understanding and integration of semantic information from different dimensions provided by T5 and CLIP.

5 Conclusions

In this report, we introduced the entire pipeline of building Hunyuan-DiT , which is a text-to-image model with the ability to understand both English and Chinese. Our report elucidates the model design, data processing and the evaluation protocol of Hunyuan-DiT . Combining these efforts from different aspects. Hunyuan-DiT reaches the top performance in Chinese-to-image generation among open-source models. We hope Hunyuan-DiT can be a useful recipe for the community to train better text-to-image models.

References

[1] Midjourney. `https://www.midjourney.com/home`.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

[4] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22669–22679, 2023.

[5] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

[6] Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Patrick Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning*, pages 4055–4075. PMLR, 2023.

[7] Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-$\backslash alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2023.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024.

[10] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023.

[11] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Joshua M Susskind, and Navdeep Jaitly. Matryoshka diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[12] Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253, 2020.

[13] 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.

[14] Minbin Huang, Yanxin Long, Xinchi Deng, Ruihang Chu, Jiangfeng Xiong, Xiaodan Liang, Hong Cheng, Qinglin Lu, and Wei Liu. Dialoggen: Multi-modal interactive dialogue system for multi-turn text-to-image generation. *arXiv preprint arXiv:2403.08857*, 2024.

[15] Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

[16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[17] Ruijun Li, Weihua Li, Yi Yang, Hanyu Wei, Jianhua Jiang, and Quan Bai. Swinv2-imagen: Hierarchical vision transformer diffusion models for text-to-image generation. *Neural Computing and Applications*, pages 1–16, 2023.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[20] Xingchao Liu, Xiwen Zhang, Jianzhu Ma, Jian Peng, and Qiang Liu. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023.

[21] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

[22] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[23] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2023.

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[25] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[27] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[28] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2021.

[29] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

[30] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[32] Chengyu Wang, Zhongjie Duan, Bingyan Liu, Xinyi Zou, Cen Chen, Kui Jia, and Jun Huang. Pai-diffusion: Constructing and serving a family of open chinese diffusion models for text-to-image synthesis on the cloud. *arXiv preprint arXiv:2309.05534*, 2023.

[33] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[34] Xiaojun Wu, Dixiang Zhang, Ruyi Gan, Junyu Lu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. Taiyi-diffusion-xl: Advancing bilingual text-to-image generation with large vision-language model support. *arXiv preprint arXiv:2401.14688*, 2024.

[35] Yanwu Xu, Yang Zhao, Zhisheng Xiao, and Tingbo Hou. Ufogen: You forward once large scale text-to-image generation via diffusion gans. *arXiv preprint arXiv:2311.09257*, 2023.

[36] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *arXiv preprint arXiv:2401.11708*, 2024.

[37] Fulong Ye, Guang Liu, Xinya Wu, and Ledell Wu. Altdiffusion: A multilingual text-to-image diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6648–6656, 2024.

[38] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. *arXiv preprint arXiv:2311.18828*, 2023.

[39] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. *arXiv preprint arXiv:2310.20550*, 2023.

A Additional Materials

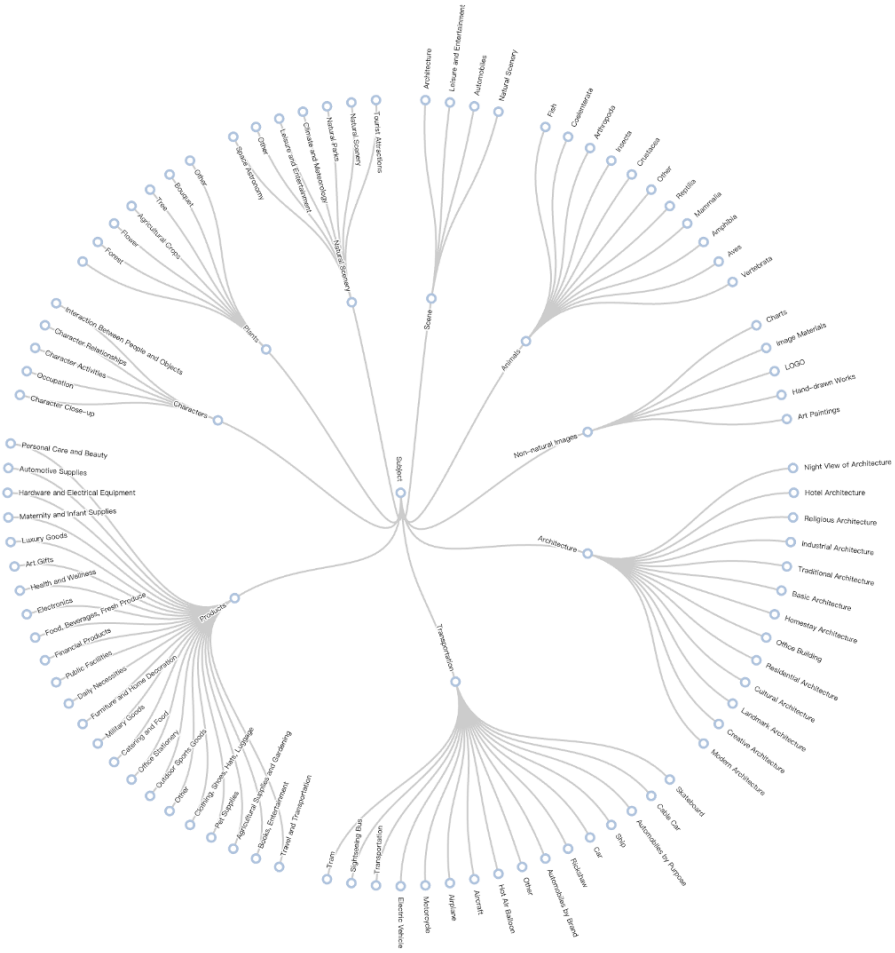


Figure 18: The hierarchy of subjects in our training data.

Categories	Weights
Functional Images	5%
Human Characters	20%
Iconic Imagery	5%
Chinese Elements	10 %
Artistic Styles	20 %
Spatial Composition	10 %
Subject and Details	20 %

Table 2: Weights of different categories in our evaluation protocol. Note that the weights are summed to 90% because we put 10% for multi-turn text-to-image generation when evaluating our own model internally. For comparison with SOTA, we only consider the cateogires in the table.

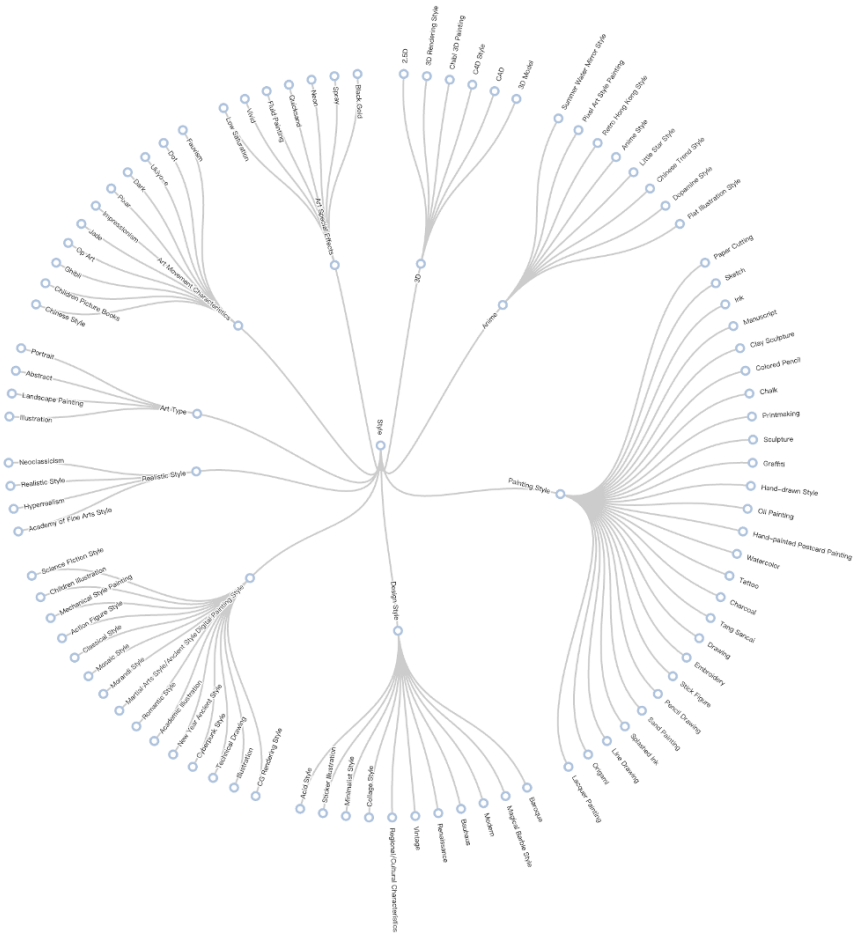


Figure 19: The hierarchy of styles in our training data.

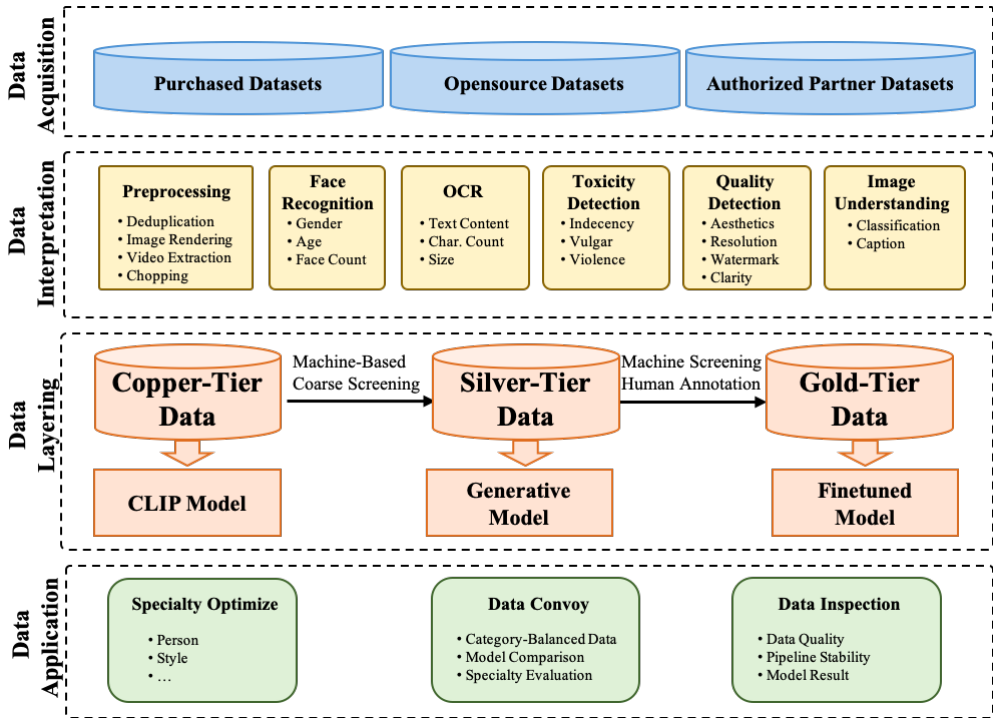


Figure 20: Illustration of our whole data pipeline.

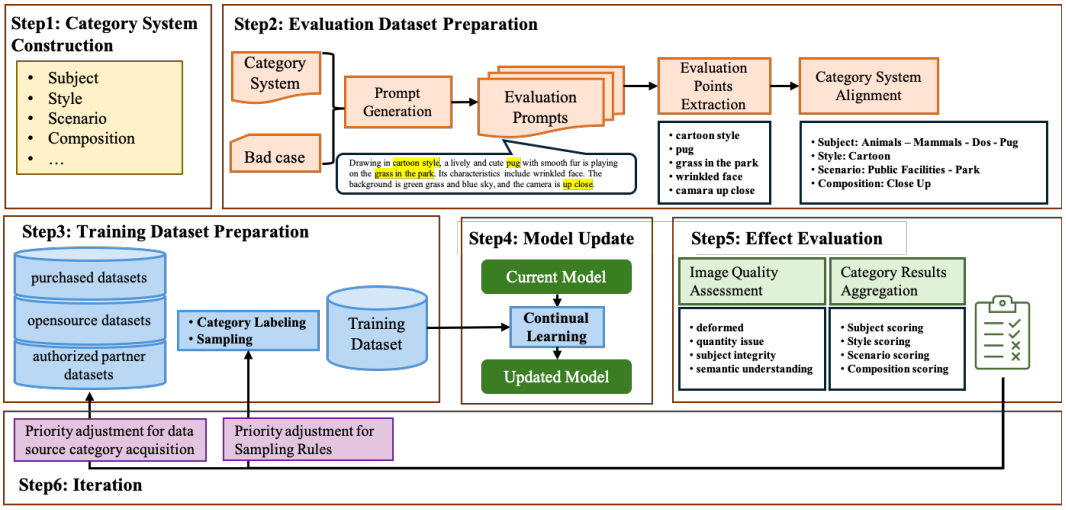


Figure 21: Illustration of our ‘data convoy’ mechanism.