Tencent Text-to-Image Generative Model

Lu Qinglin

2024.3.21

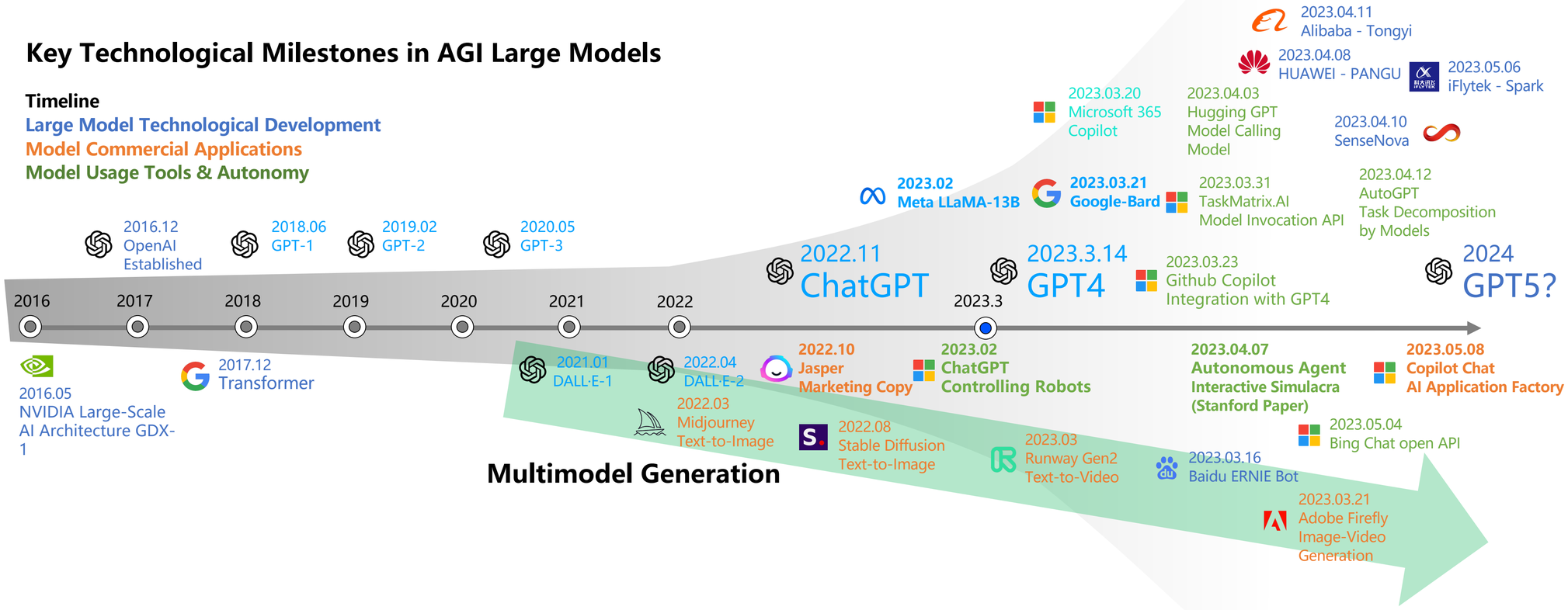
Tencent Hunyuan

Contents

Tencent Hunyuan

Background

AGI Revolution: **Multimodal Generation at the Core**



Tencent Hunyuan

Tencent: Diverse Products and Extensive Text-to-Image Applications

1. Advertising Scenarios

Generation of product advertisements

2. Gaming Scenarios

Generation of game elements

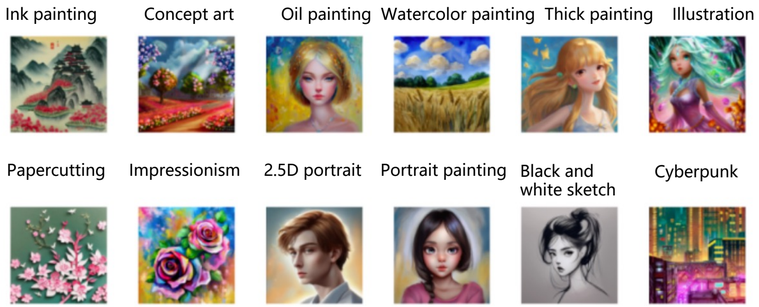
Creation of game characters

3. Content Scenarios

Domineering CEO stories

Generation of novel illustrations

4.Cloud Service Scenarios



Tencent Hunyuan

Tencent Text-to-Image Foundation Model

Tencent Hunyuan

Three major challenges of the Text-to-Image model:
Semantics, Rationality, Aesthetics

Challenges

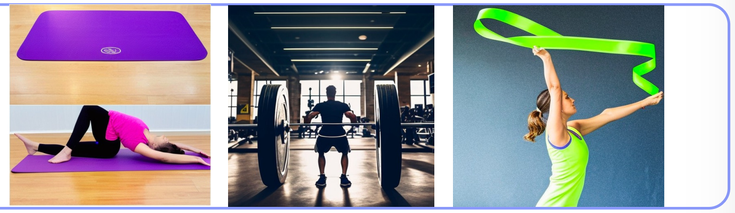
1.Difficult to accurately follow the text instruction

手持青龙偃月刀的关公

在黄昏的美景背景下，一架黄色的飞机正在机场跑道上起飞

三个女孩手拿饮料站在室内的饮料机旁

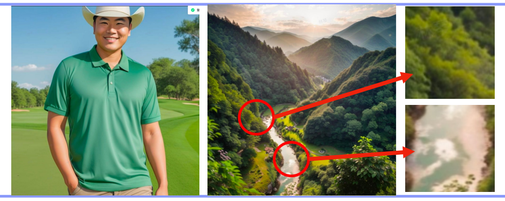
2. Prone to human deformities and composition



3.Insufficient drawing details in human and scenery

insufficient detail in character

poor attention to detail



Goal

Creating the industry-leading Tencent Hunyuan Foundation Model for Text-to-Image Generation

Solutions

「Accurate semantics」

- bilingual fine-grained CLIP model, ensuring a thorough understanding of both subject and attributes.
- The prior model connects image-text features, enhancing semantic expression.

「Rational Structure」

- transformers-based framework，strengthening the rational expression of content structure
- Introduce prior human body structure information to correct the generation of human body structure

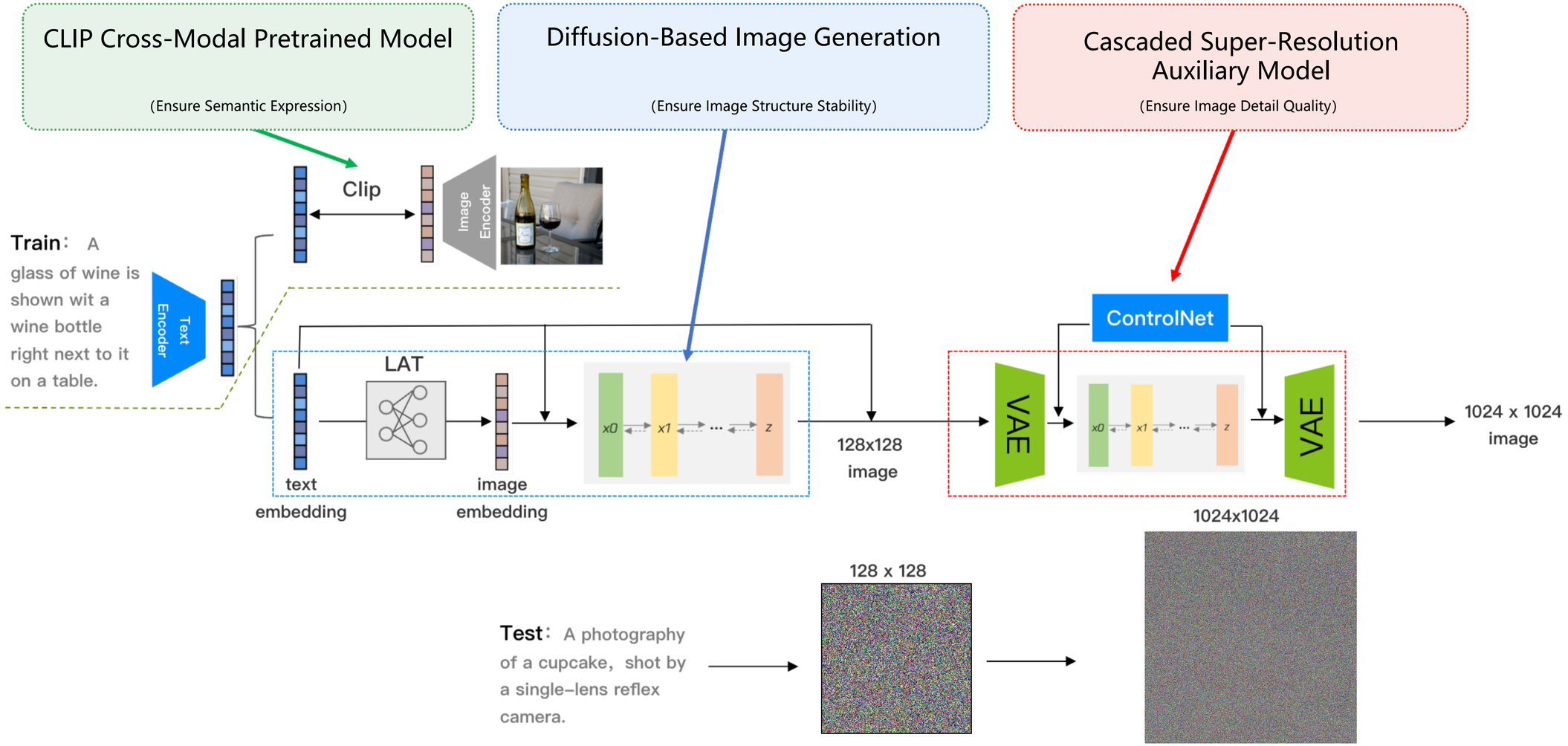
「Aesthetic Details」

Construct amodel fusion scheme to fully demonstrate the details of characters and scenes in various dimensions of the model

Tencent Hunyuan

The structure of the Hunyuan Text-to-Image model:

1. Ensuring the semantic precision 2. structural reasonableness 3. detail quality of the generated images comprehensively.



[Semantic Understanding Expression]
Hunyuan CLIP supports bilingual Chinese-English and fine-grained expression

Difficulties and Challenges

1. Depends on translation to implement Chinese drawing capabilities, it's difficult to accurately generate in a Chinese context.

女娲补天
Nuwa mends the sky

2. The perceptual capabilities of the CLIP image-text alignment model are limited, impacting the text-to-image model's ability to generate fine-grained details.

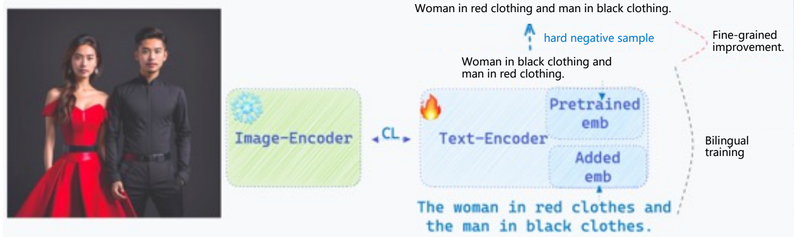
Woman in black clothing and man in red clothing (fine-grained attribute error generation).

Solution Approach

① Simultaneously model Chinese and English to attain bilingual comprehension.

② Improve the distinctiveness of textual features in fine-grained attributes.

Technical Solution



Innovation 1

Enhance the tokenizer and integrate it with Chinese-English image-text pair training to boost bilingual comprehension and increase encoding efficiency.

Innovation 2

Improve the fine-grained differentiation capabilities of the CLIP model by employing contrastive learning and generating challenging negative samples.

Results

English-Only	Company	Params	ImageNet1K-CN	ImageNet1K-EN
(2021) CLIP	OpenAI	428M	-	76.6
(2022) OpenCLIP	-	1.0B	-	**78.0**
Bilingual(Chinese+English)				
(2022) Taiyi-CLIP	IDEA	958M	54.4	-
(2022) CN-CLIP	Alibaba	958M	58.8	32.3
(2022) AltCLIP	BAAI	864M	59.6	74.5
(2023) **HunYuanCLIP**	**Tencent**	984M	**66.4**	71.2

Table 3: Experimental results ImageNet1K-CN and ImageNet1k-EN

Across multiple datasets, including ImageNet, the performance metrics for Chinese-specific models surpass those of bilingual counterparts, such as AltCLIP and CN-CLIP.

Before Optimization

After Optimization

✨ A person wearing a brown coat holding a blue coffee cup.

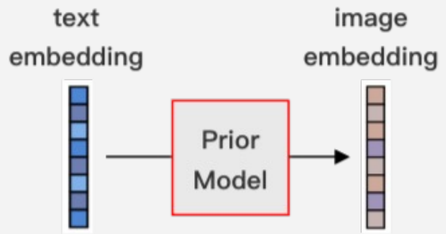
Tencent Hunyuan

[Semantic Understanding and Expression]

Lightweight prior for cross-modal representation, using images to enrich textual semantic information

Difficulties and Challenges

Cross-domain generation of text & images, how to link text features with image features? [Prior Model]



Solution Approach

① **Using diffusion models as prior models (Dalle2) involves high training complexity, excessive computational demands, and elevated optimization costs.**

② **Not using prior models (Imagen) and directly** learning **across domains can be too aggressive, leading to high training data costs.**

Solution: A lightweight prior (Local Affine Transform) model creates a bridge across modalities, trading small computational loads for stable generation results.

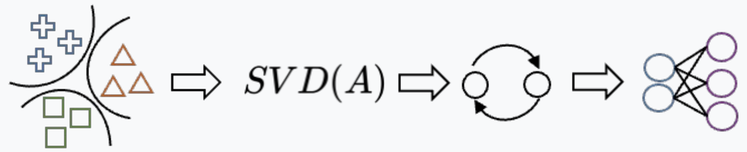
Step 1: Cluster first, then regress. Maintain local linearity throughout the change process.

Step 2: Perform low-rank decomposition on the matrix A calculated after clustering, making the transformation matrix present in a low-rank form to enhance generalizability.

$$A = USV^{\top}$$

$$\begin{cases} k = 1, 2 \cdots \\ U_k = \left(YX^{\top}V_{k-1}\right)\left(V_{k-1}^{\top}XX^{\top}V_{k-1}\right)^{-1} & d_H \times k \\ V_k = \left(XX^{\top}\right)^{-1}\left(XY^{\top}U_k\right)\left(U_k^{\top}U_k\right)^{-1} & d_L \times k \end{cases}$$

Step 3: Optimize using the Alternating Least Squares method until convergence is achieved, obtaining matrices U and V for feature transformation.



Cluster | Regress and Factorize | Alternating Least Squares | Transformation matrix

Results

Advantageous Result 1:

Compared to DallE2's prior, with only 0.44% **of the parameters, incorporating both image and text embeddings into the geneartion model, FID can be reduced by 1.3**

	FID
DallE2	10.4
LAT Image Emb	12.1
LAT Image-Text Emb	9.1

Advantageous Result 2:
Plug and play,
with strong extensibility.

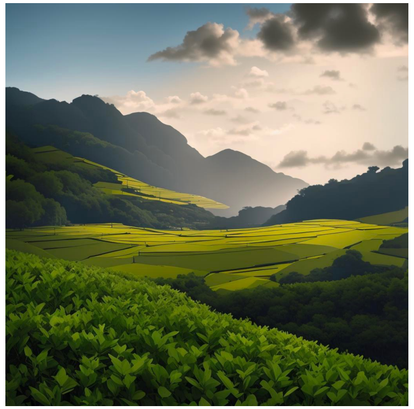
Tencent Hunyuan

[Semantic Understanding]
Generated images can more accurately depict semantics

Before optimization

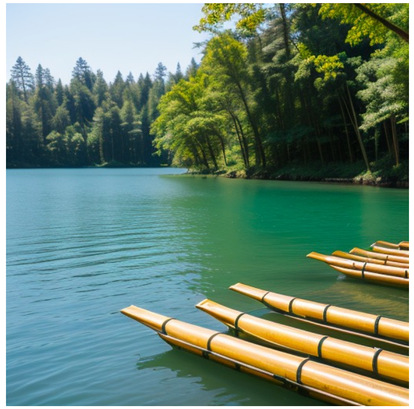
After optimization

[Clouds in the heart]



[A target on an apple]

[Fishermen sing at dusk]



Tencent Hunyuan

[Rational Generation Structure]
Improving the Rationality of Layout and Structure via Diffusion with Transformer

Difficulties and Challenges

How to enhance the rationality of content layout and structure in generated images?

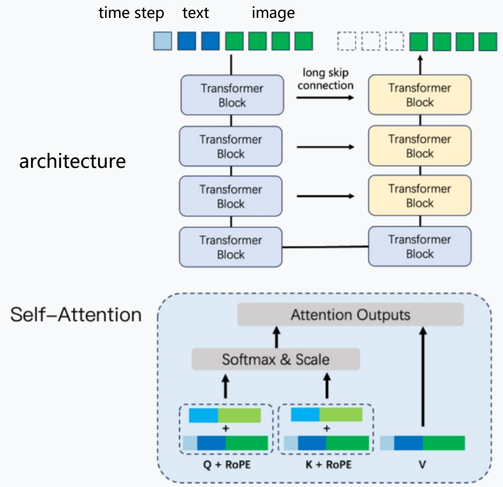
This is a painting depicting the famous Parisian landmark, the Eiffel Tower, centered against a brightly white brick wall background. The artwork is vintage and rich in French landmark culture, exuding a strong nostalgic atmosphere.

Solution Approach

① **In the latent space, employing a fully Transformer architecture for the diffusion model enables bidirectional attention computation between text and images, achieving a thorough alignment between text and image spaces.**

② **The use of one-dimensional and two-dimensional rotary position encoding for text and images respectively enhances the model's spatial awareness.**

Technical Approach



Innovation Point 1

Pure Transformer architecture for the diffusion model, unifying image and text tokens to implement bidirectional attention. This results in more precise text-to-image control.

Innovation Point 2

Implementing rotary position encoding within the Transformer enhances the two-dimensional spatial awareness of images. The resulting images have more rational structures with lower distortion rates.

Results

The rate of good cases increased from 55% to 67%

✨ **Circular food placed on a heart-shaped wood**

✨ **H**eart-shaped food placed on a circular wood

Tencent Hunyuan

[Reasonableness of generated structure]
Original twin network EnhanceNet, optimizes the reasonableness of human body structure

Difficulties and Challenges

Challenge 1: Generating a human body without guaranteeing structural correctness

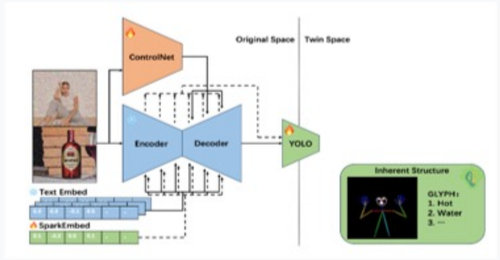
Challenge 2:Deformed hands in generation

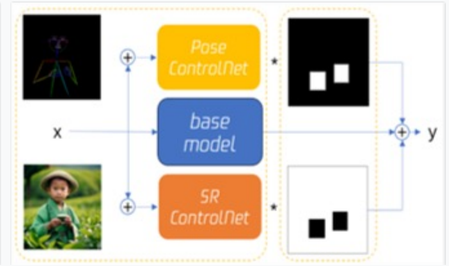
Solution Approach

① Human skeletal information has a certain prior structure, and this prior information can be introduced into the generation process.

② There are differences in the level of detail between human body structure and hand structure, which need to be introduced from different stages.

Technical solution: Repairing with the skeleton prior based on ControlNet in both the generative model and the super-resolution model.





Innovation Point 1

Based on the Siamese network, we construct EnhancedNet to correct deformities in human limbs.

Innovation Point 2

Based on a multi-control approach, synchronously repair **local deformities of fingers**.

Results

Input	mAP50 (↑)				mAP50-95 (↑)			
	200	400	600	800	200	400	600	800
TextEmbed	0.792	0.767	0.676	0.376	0.649	0.612	0.458	0.173
SparkEmbed	0.81	0.781	0.679	0.373	0.689	0.636	0.47	0.178
PredZ	0.81	0.788	0.695	0.387	0.681	0.652	0.494	0.181
Spark + PredZ	**0.824**	**0.797**	**0.696**	**0.389**	**0.706**	**0.668**	**0.494**	**0.186**

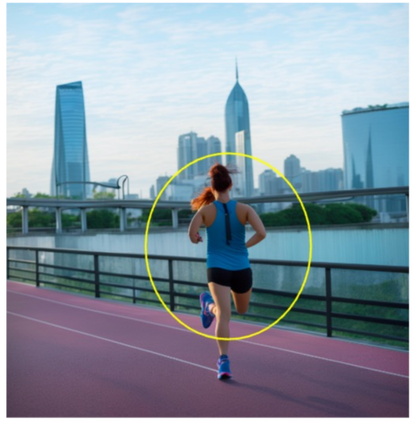
Tencent Hunyuan

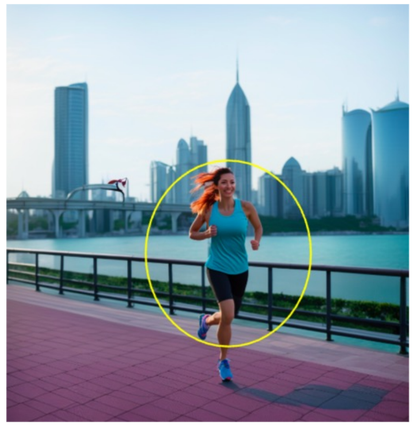
[Reasonable Generation Structure]

Original Siamese Network EnhanceNet, optimizing human body structure generation.

Before optimization

After optimization



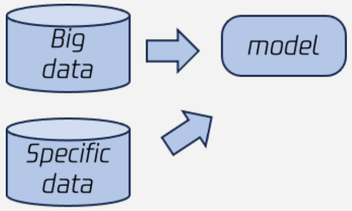


Tencent Hunyuan

[The details and texture of the image]
Utilizing a fusion model-based method to enhance the texture of the generated images.

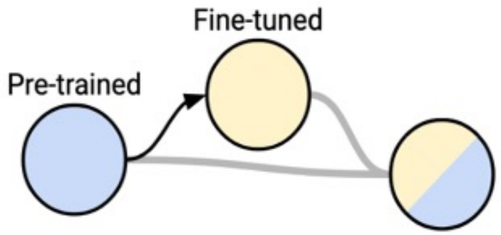
Challenges

The volume of training data reaches the level of hundreds of millions, while the specialized data volume is only at the level of tens of thousands. The joint fine-tuning of pre-trained models for specialized data modeling is weak, and the cycle is long

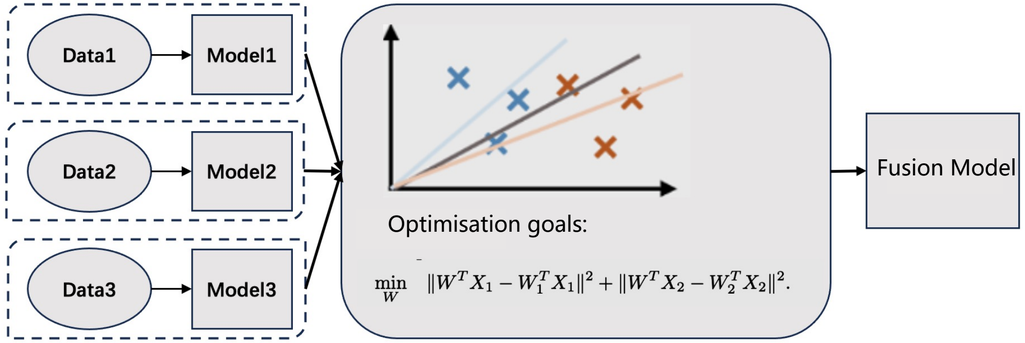


Our Approach

Fine-tuning specialized data on top of the pre-trained model, then integrating it with the pre-trained model for knowledge **transfer.**



Technical programme：MIF (model iterative fusion) improves the efficiency and effectiveness of training



key point 1

Turning model fusion into an optimisation problem, the new model fits the performance of all models on their respective data, improving optimisation efficiency

key point 2

Iteratively update the parameters using a low-rank decomposition to update only the most important parameters to avoid over-fitting and improve model generation.

Results

portrait model (hair, wrinkles), 30% increase

scene models (grass, trees, ripples), 25% increase

Training time from 5 days to 1 day

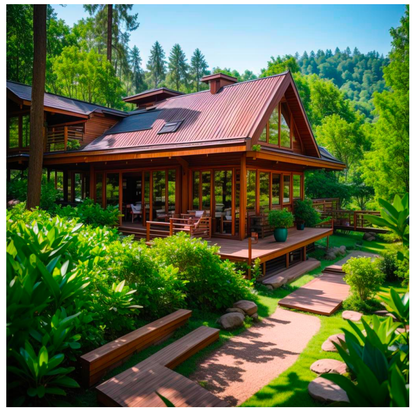
Tencent Hunyuan

[Image Detail Texture]

The generated images can depict texture and quality in greater detail.

Before optimization

After optimization



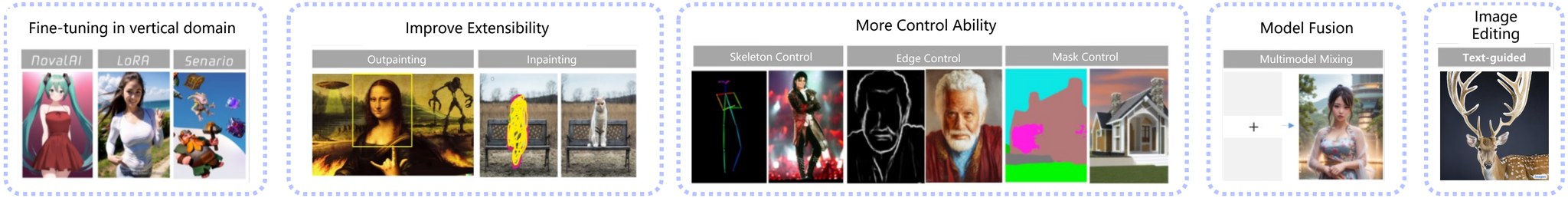
Tencent Hunyuan

Text-to-Image Adapters

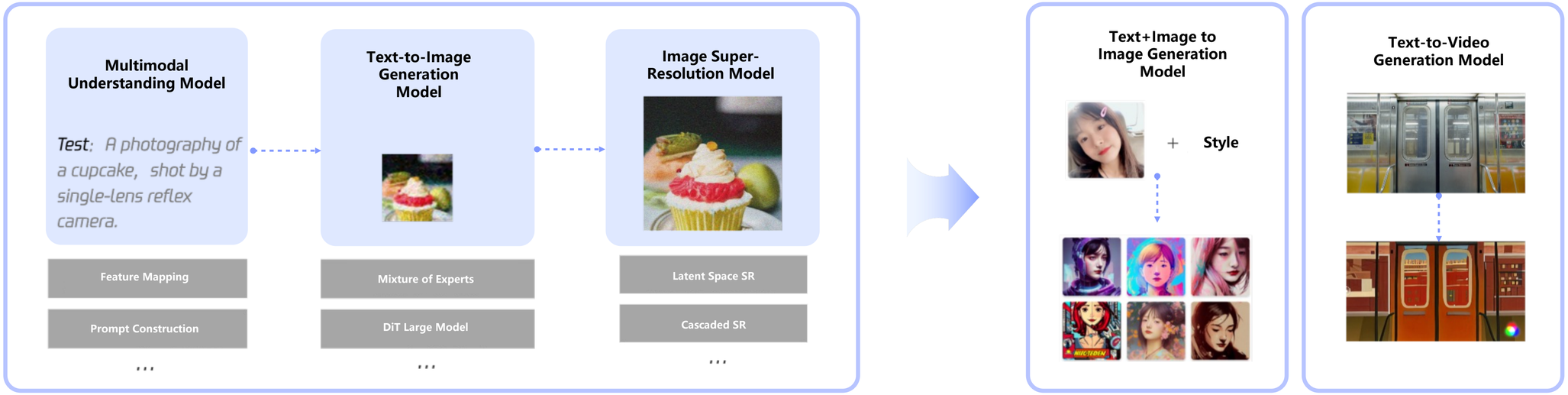
Tencent Hunyuan

Tencent HunYuan Text-to-Image Model Panorama

Task-Specific Light Adapters



Foundation Large Model



Training Data

Large-scale text-image pairs

Tencent Hunyuan

1. Game Asset Generation

2D ancient style, a woman in ancient style, wearing white Hanfu

2. Product Background Generation

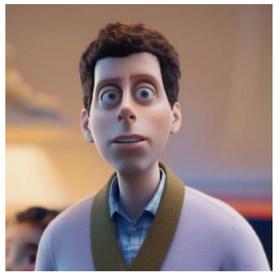
Product

3. image2image translation

Image from Customer

4. PhotoMaker: Portrait Editing

Portrait



Tencent Hunyuan

1. Game Asset Creation: A variety of 2D and 3D art styles

2D art style: The brushwork is like a finely drawn flat animation, as if it's a hand-drawn or mouse-drawn work by a high-level artist.

3D art style offers a sense of depth, strong interplay of light and shadow, and visual effects akin to those produced by 3D modeling software .

Anime

Japanese Anime | Chibi | ...

Chinese-Style

Chinese Ancient | Ink and Wash | ...

Sci-Fi

Science Fiction | Military | ...

Western Fantasy

Fantasy characters | Magical landscape | ...

Tencent Hunyuan

1. Game Asset Creation

2D art style: Currently, it can generate modern, ancient style, ink wash, chibi characters, vector, and Western fantasy style art materials.

Japanese Anime

Anime, a photographer holding a camera, dressed in casual clothes, situated in the midst of nature.

Chinese-Style

Chibi, an anime girl, with little cat ears, wearing a school uniform, standing inside the campus.

Tencent Hunyuan

1. Game Asset Creation

2D art style: Currently, it can generate modern, ancient style, ink wash, chibi characters, vector, and Western fantasy style art materials.

Chinese Ancient Character

The ancient style, a handsome general, dressed in battle robes, stands on the battlefield.

Chinese Ink Wash Landscape

Ink wash style, ancient courtyard, lakes and trees.

Tencent Hunyuan

1. Game Asset Creation

3D art style: Currently, we can generate materials in various art styles, including modern, ancient, Western fantasy, cyberpunk, futuristic Sci-Fi, military gaming, racing gaming, and chibi-style animals.

Sci-Fi Character

A military game character, a female special forces soldier, amidst the war-torn ruins.

Military Game Character

A science fiction character, an Asian woman, dressed in a white dress, standing on the city street.

Tencent Hunyuan

1. Game Asset Creation

3D art style: Currently, we can generate materials in various art styles, including modern, ancient, Western fantasy, cyberpunk, futuristic Sci-Fi, military gaming, racing gaming, and chibi-style animals.

Western Fantasy Character

CG art style, A stunningly beautiful girl dressed in medieval clothing, wearing a headpiece and jewelry, and donning a flowing skirt.

Western Fantasy Landscape

A Western fantasy scene, featuring a magnificent Baroque palace, shimmering with gold and splendor

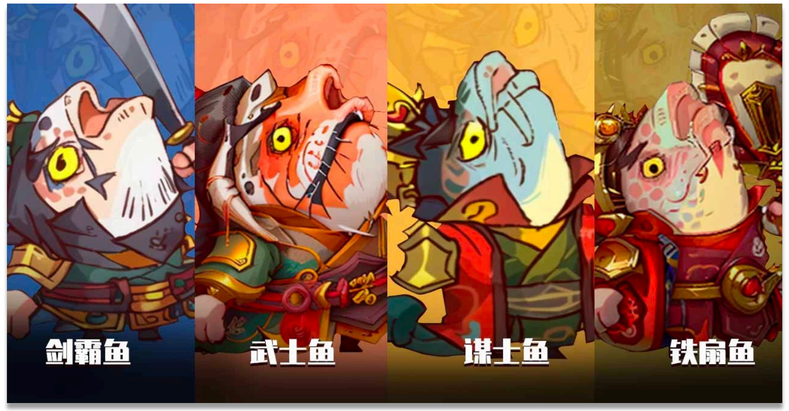
Tencent Hunyuan

1. Game Asset Creation

Fine-tuning with LoRA: A stylized small model based on around 10 images, meeting the generation demands of specific images in the game industry.

DaoTian Lu

King of Salted Fish



Xundao Daqian

Chest Knight

Tencent Hunyuan

2. Product Background Generation

Support controlled generation of rich and realistic backgrounds for product subjects.

Clothing, shoes, and bags

Skincare and Beauty

Soft Drinks and Food

Consumer Eletronics and 3C

Personsal Care and Household Cleaning

.....

Covering various product categories in e-commerce consumer goods
& extracting mainstream professional commercial photography backgrounds

AIGC ✚

Bathroom Countertop

Dining Table Surface

Forest Environment

Autumn Leaves

Rainbow gradient

.....

Tencent Hunyuan

2. Product Background Generation

Health & Personal Care Products

E-commerce Customer: Specified Product Background Generation

"Placed in an environment filled with fresh flowers, a minimalist style, and a refreshing, natural atmosphere—advertising photography"

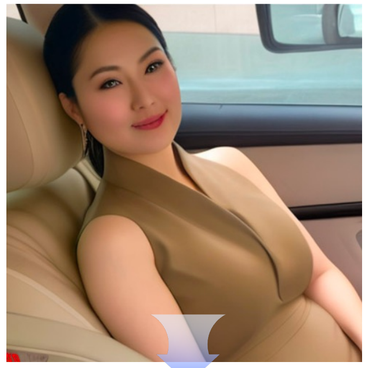
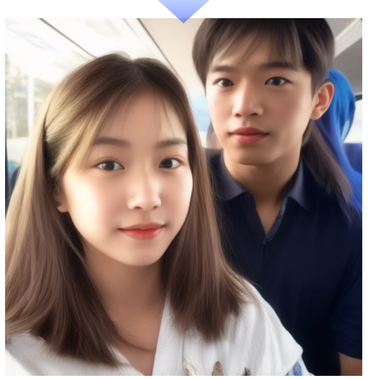
"Placed in an environment filled with marble, a minimalist style"

Product

Tencent Hunyuan

3. Image2image translation

Health & Personal Care Products

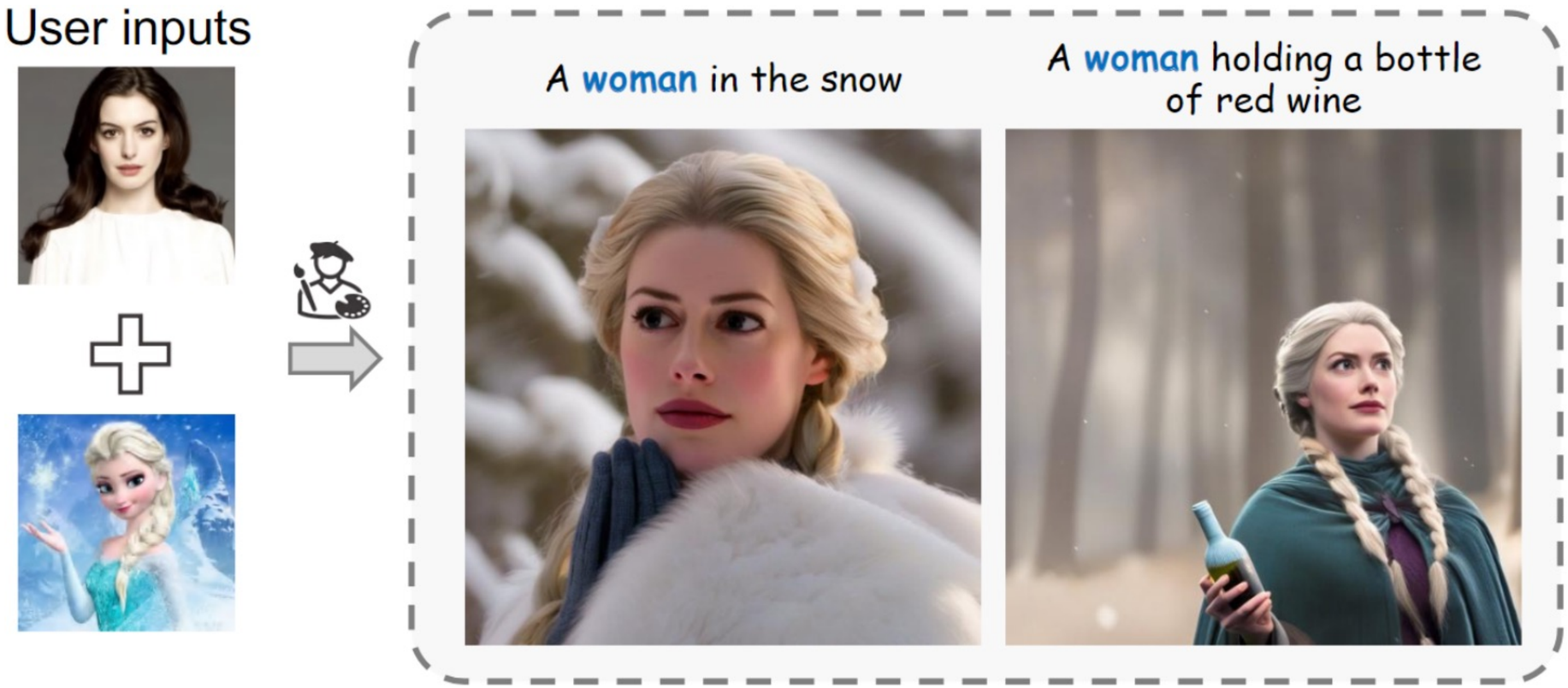
	Keep Actions	Keep Temperament	Keep Relationship
Original Image			
Image to Image			

Tencent Hunyuan

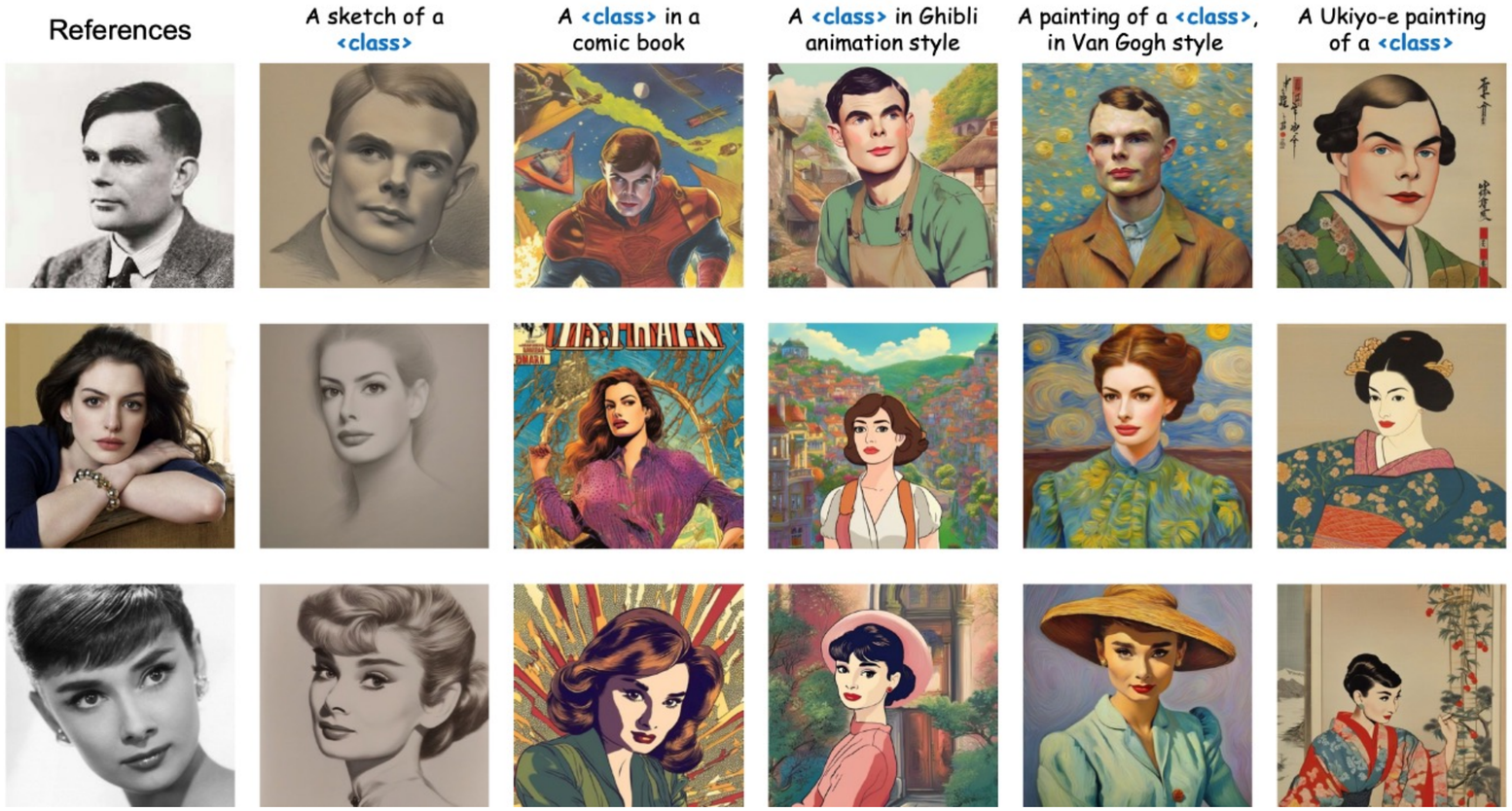
4. PhotoMaker: Attribute Manipulation

A novel training-free portrait editing method.

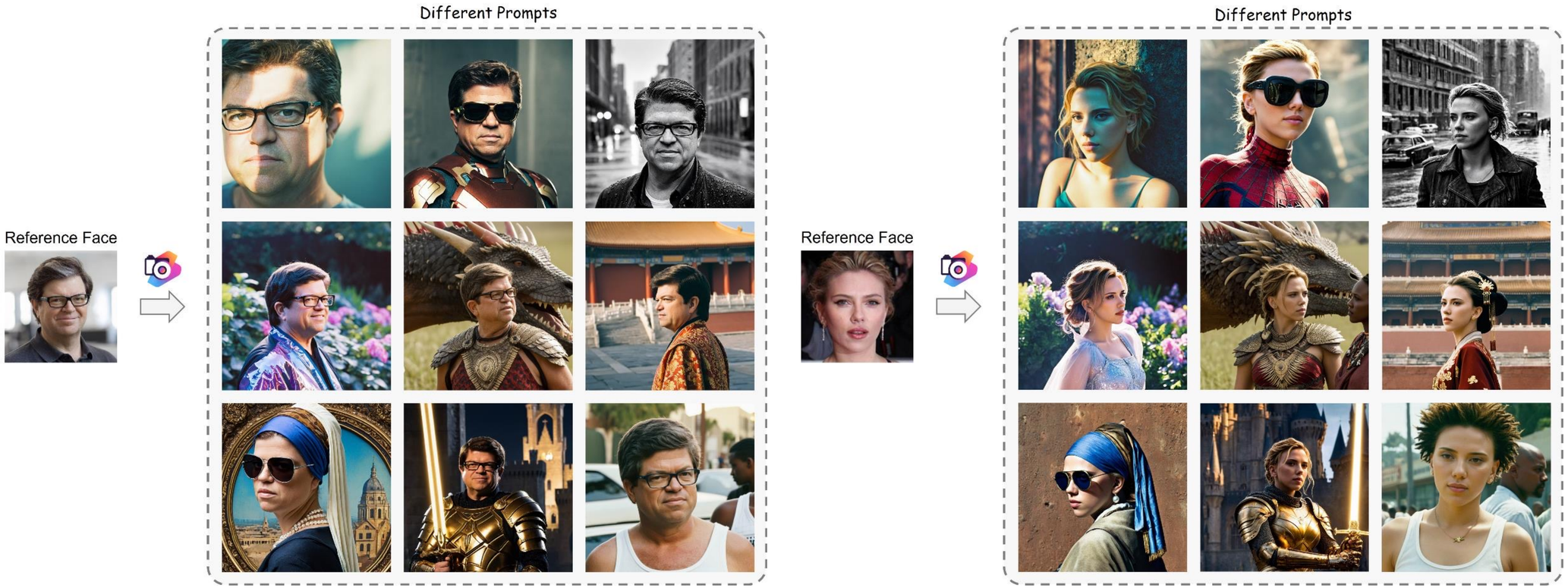
4. PhotoMaker: Identity Mixing



4. PhotoMaker: Stylization



4. PhotoMaker: More examples in the wild



4. PhotoMaker: More examples in the wild

Applications

Tencent Hunyuan

Ads application: material creation, product synthesis, and game material production

1. material creation

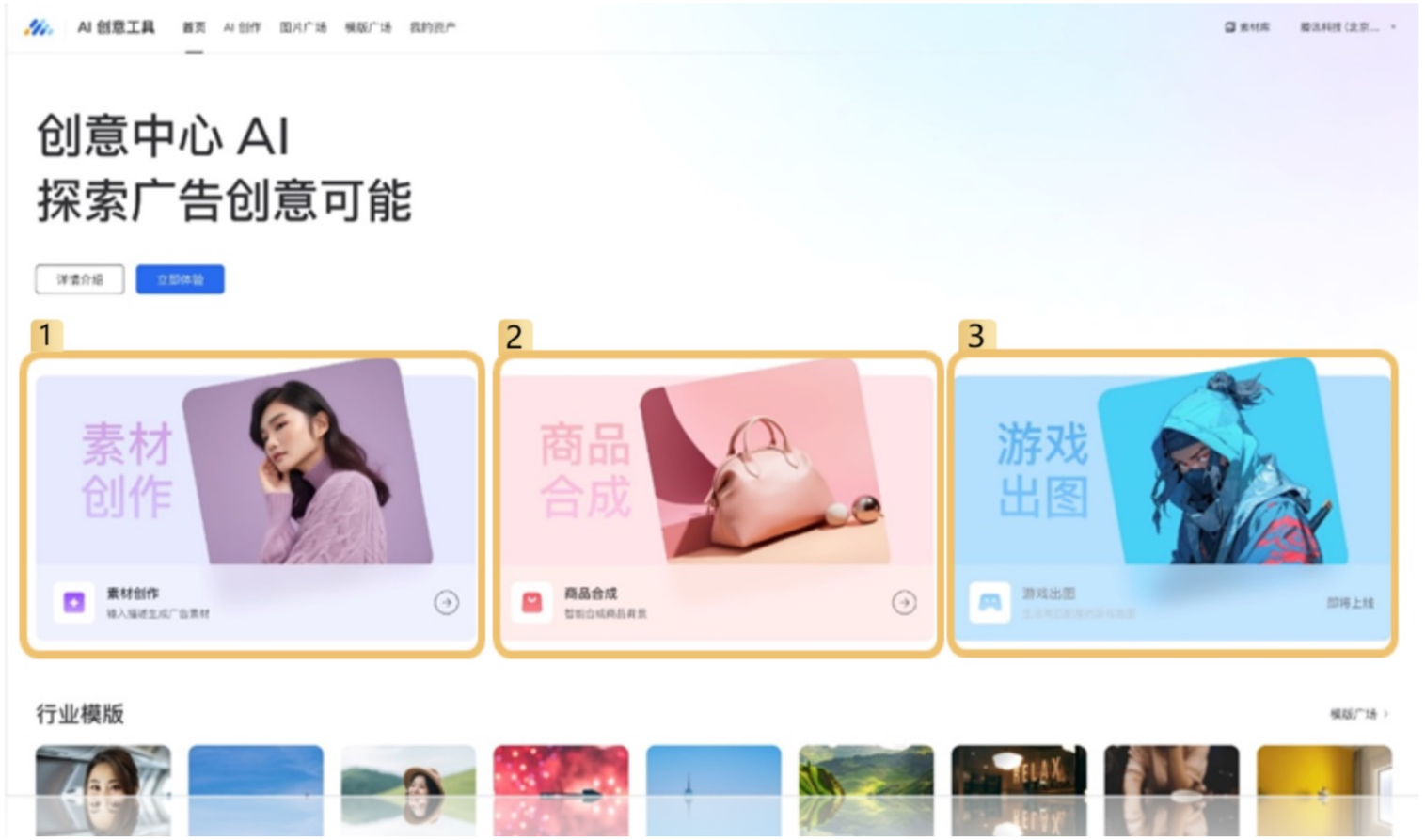
Automatically output diverse ideas by prompt

2. product synthesis

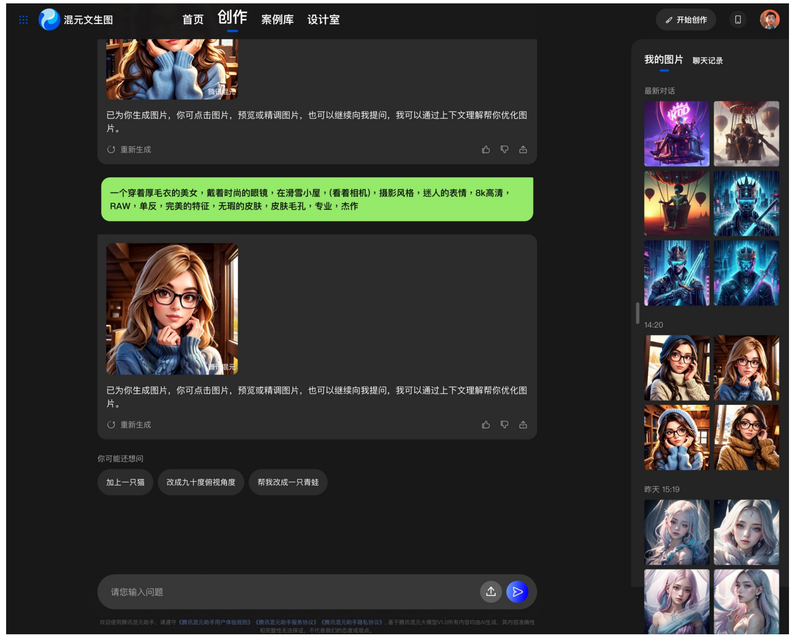
Connect to customer product library and support controllable product entities to generate real backgrounds

3. game material production

Generate AI advertising creative ideas based on the advertiser's specific style and IP role



Creation module: It can provide multiple rounds of image and text, multi-modal dialogue, tool-based image editing and other capabilities.



Multi-round dialogue with images and texts:

Based on the natural language input by the user, images related to the description are generated and prompt words are mapped, which can support multiple rounds of dialogue to supplement the description.

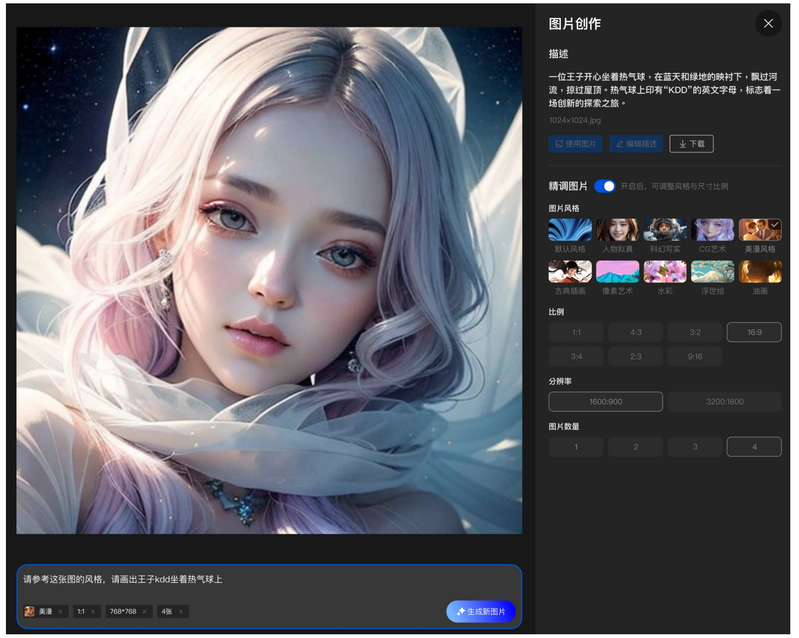


Image fine-tuning:

You can optimize and adjust the image style, size, resolution, etc. of the generated images.

Tencent Hunyuan

Laboratory module: supports more advanced graphics-generating gameplay, explores more capabilities of the Hunyuan graphics-generating model, combines B-side applications/C-side experience, and packages it as product functions.

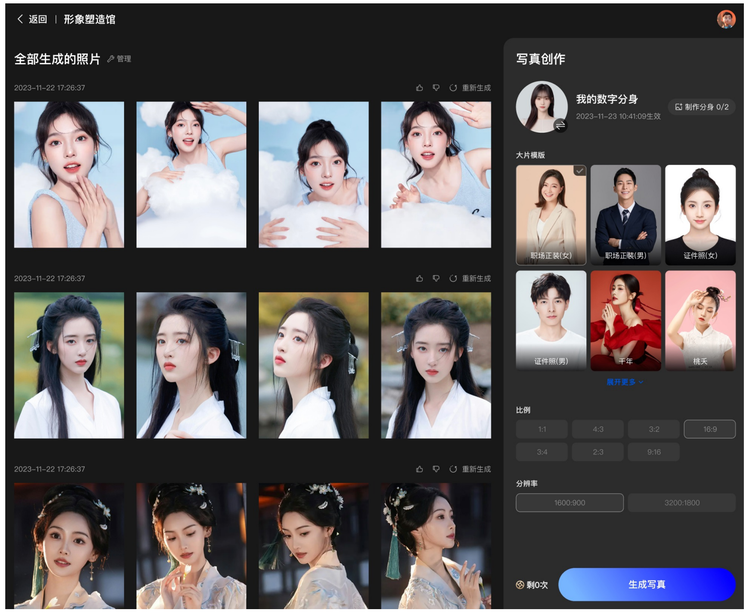
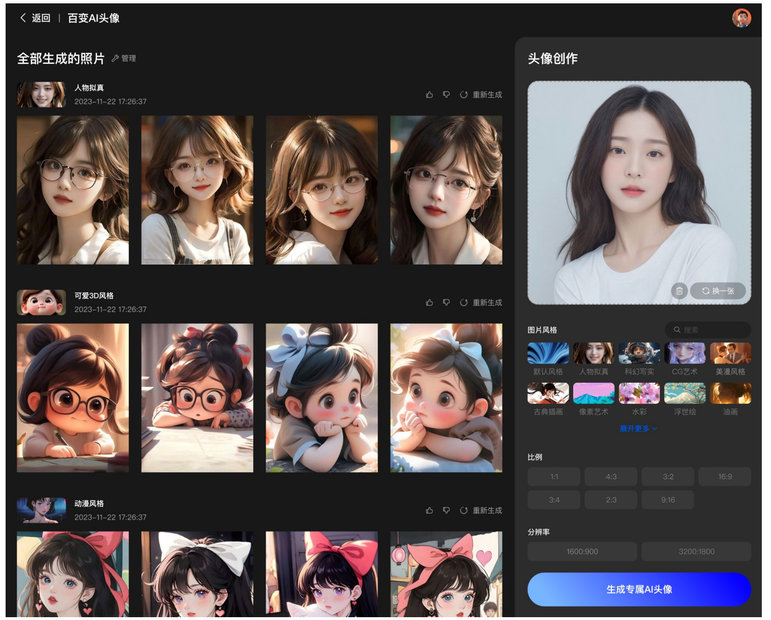


Photo Studio:

Based on about 10 training data uploaded by users, digital avatars are generated, and combined with rich photo templates, various photos are generated for users.



Variety of AI avatars:

After uploading photos, the model provides 50+ image styles for users to generate avatars of various styles.

Tencent Hunyuan

THANKS

Tencent HunYuan
Text-to-Image | Qinglinlu

NVIDIA GTC

3 月 22 日 12:30
生成式 AI

AI 时代的大会

2024 年 3 月 18 — 21 日

美国加州圣何塞及线上

芦清林
腾讯

生成式 AI 大模型
— 文字生成图像

Tencent Hunyuan