

# Supplement material of PIXART- $\Sigma$ : Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation

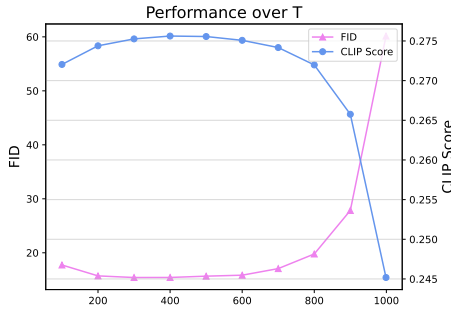
Junsong Chen<sup>1,2,3\*</sup>, Chongjian Ge<sup>1,3\*</sup>, Enze Xie<sup>1\*†</sup>, Yue Wu<sup>1\*</sup>, Lewei Yao<sup>1,4</sup>,  
Xiaozhe Ren<sup>1</sup>, Zhongdao Wang<sup>1</sup>, Ping Luo<sup>3</sup>, Huchuan Lu<sup>2</sup>, and Zhenguo Li<sup>1</sup>

<sup>1</sup> Huawei Noah’s Ark Lab <sup>2</sup> Dalian University of Technology <sup>3</sup> HKU <sup>4</sup> HKUST  
Project Page: <https://pixart-alpha.github.io/PixArt-sigma-project/>

## 1 Appendix

### 1.1 Extension: Inference Acceleration (PIXART + DMD).

**Experiments.** To expedite the inference, we integrate PIXART with DMD [4], a one-step inference technique achieved through distribution matching distillation. We train a one-step Generator  $G_\theta$ , and the generated image is denoted as  $x_0 = G_\theta(\bar{T}, \text{text})$ . Initially, we set  $\bar{T}$  to the same value as the denoising timesteps during training, that is  $\bar{T} = 999$  following [4]. However, we observed undesired results and investigated the most suitable value for  $\bar{T}$  as shown in Fig. 1. Surprisingly, the optimal  $\bar{T}$  was found to be 400 rather than 999. This deviation arises from the fact that a smaller  $T$  enhances the model’s confidence in predicting the noise of  $G_\theta$ . However, this principle is effective only within a certain range. If  $T$  becomes too small, the scenario significantly deviates from the training setting of the base model. Therefore, there exists a trade-off in selecting the  $T$  value. Finally, we compare our method quantitatively as well as qualitatively with PIXART + LCM [1] in Tab. 1 and Fig. 2.



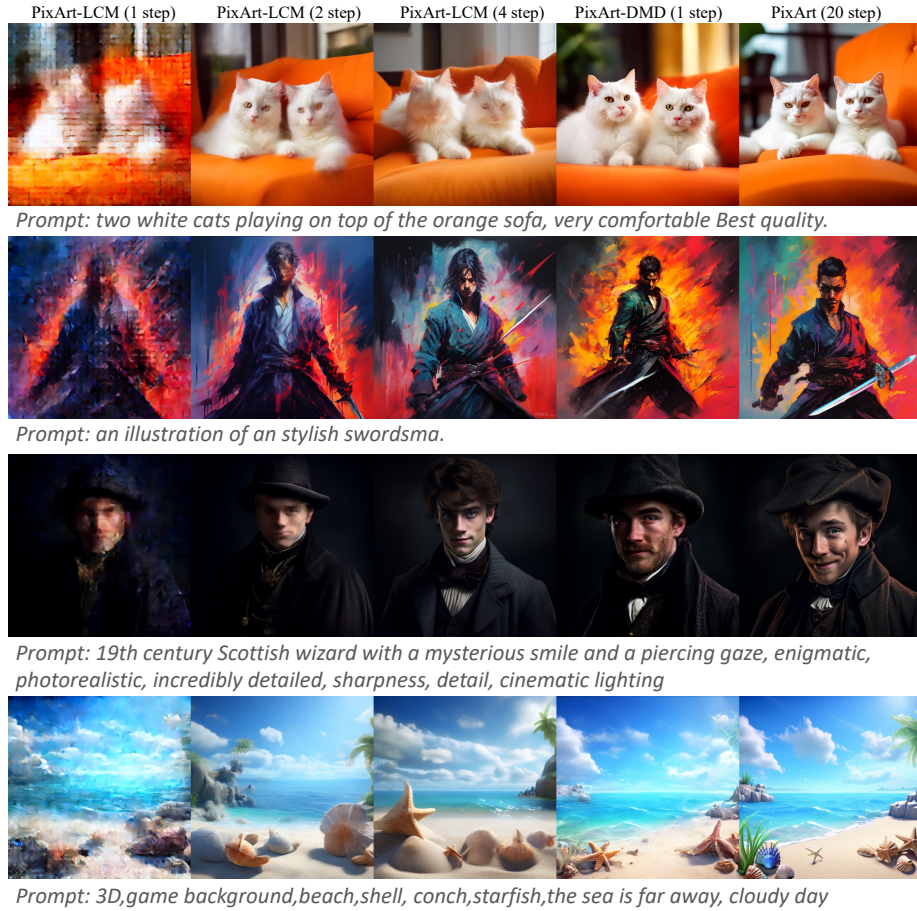
**Fig. 1:** Base model + DMD performance over  $\bar{T}$ .

**Table 1: Comparison of PIXART + DMD performance compared to PIXART + LCM.** These experiments are conducted on 512x512 resolution with a batch size of 1.

Method	FID↓	CLIP↑	Speed↓
PIXART + LCM (1 step)	108.66	0.2247	0.11s
PIXART + LCM (2 step)	17.95	0.2736	0.16s
PIXART + LCM (4 step)	13.06	0.2797	0.26s
PIXART + DMD (1 step)	13.35	0.2788	0.11s
Teacher model (20 steps)	9.273	0.2863	1.44s

### 1.2 Training Details of PIXART- $\Sigma$

In Tab. 2, we provide detailed information on each training stage of PIXART- $\Sigma$ , including image resolution, the total volume of training samples, the number of



**Fig. 2:** Visual comparison of Base model + DMD performance *vs.* Base model + LCM model.

training steps, batch size, learning rate, and the computing time measured in GPU days. Utilizing the Internal- $\Sigma$  dataset and integrating a more advanced VAE, our proposed method quickly adapts to the new VAE, requiring a mere 5 V100 GPU days. Subsequently, we achieve remarkable text-image alignment with just 50 V100 GPU days, showcasing a substantial improvement over PIXART- $\alpha$  at minimal additional training expense, thanks to the proposed "weak-to-strong" strategy, which proves highly efficient.

Notably, applying KV token compression stands out as a significant efficiency booster, markedly reducing the training duration. For example, when fine-tuning from 512px to 1024px and incorporating KV compression, the training time reduces dramatically from 50 V100 GPU days to a mere 20 V100 GPU days. Likewise, for resolutions of 2K and 4K, the training time diminishes from 20 to 14 A800 GPU days and from 25 to 20 A800 GPU days, respectively. This

underscores the remarkable efficacy of KV token compression in augmenting training efficiency.

**Table 2:** We report detailed information about each training stage of PIXART- $\Sigma$ . Note that Internal- $\Sigma$  dataset here includes 33M internal data. The count of GPU days excludes the time for VAE feature extraction and T5 text feature extraction, as we offline prepare both features in advance so that they are not part of the training process and contribute no extra time to it.

Stage	Image Resolution	#Images	Training Steps	Batch Size	Learning Rate	GPU days
VAE adaption	256×256	33M	8K	64×16	$2 \times 10^{-5}$	5 V100
Better Text-Image align	256×256	33M	80K	64×16	$2 \times 10^{-5}$	50 V100
Higher aesthetics	512×512	18M	10K	32×32	$2 \times 10^{-5}$	30 V100
Higher aesthetics	1024×1024	18M	5K	12×32	$1 \times 10^{-5}$	50 V100
KV token compression	1024×1024	18M	5K	12×16	$1 \times 10^{-5}$	20 V100
Higher aesthetics	2K×2K	300K	4K	4×8	$2 \times 10^{-5}$	20 A800
KV token compression	2K×2K	300K	4K	4×8	$2 \times 10^{-5}$	14 A800
Higher aesthetics	4K×4K	100K	2K	4×8	$2 \times 10^{-5}$	25 A800
KV token compression	4K×4K	100K	2K	4×8	$2 \times 10^{-5}$	20 A800

### 1.3 Detailed Settings of the Ablation Studies

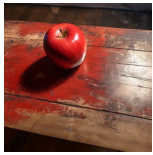
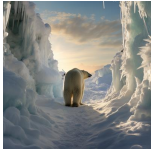
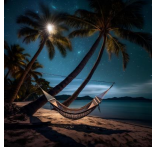

In this subsection, we describe the additional experimental setup for the ablation studies. For each study, we evaluated the Fréchet Inception Distance (FID) by comparing images generated by our PIXART- $\Sigma$  using a set of 30,000 curated prompts against those from the High-Quality Evaluation Dataset. Additionally, we assessed the Clip-Scores by comparing the model-generated images to the original 30,000 prompts. Our findings suggest that the number of images used for testing can influence the FID scores; for instance, a dataset comprising 10,000 images typically yields higher FID scores. Besides, for the conducted experiments on compression positions and operators, we test the respective models with images generated at a resolution of 512px.

### 1.4 Samples of High-Quality Evaluation Dataset

We observe the evaluations performed on the MSCOCO dataset could not adequately fully capture a model’s proficiency in aesthetics and text-image alignment [2, 3]. Thus, we propose an evaluation dataset that consisting of 30,000 high-quality, aesthetically pleasing text-image pairs for a more thorough assessment of a model’s ability to generate visually appealing images. The visualizations in Fig. 3 exemplify the dataset’s superior aesthetic quality and the precise alignment between textual descriptions and visual content.

### 1.5 FID Comparisons with Open-source Models

We conducted comparative analyses of open-source models using FID and CLIP-Scores on our curated dataset. The results, presented in Tab. 3, reveal that the

Image	Prompt	Image	Prompt
	A red apple sitting on a wooden table, remote control aerial photography.		A photographic work capturing a polar bear walking through icy and snowy terrain.
	A serene beach with palm trees, turquoise water, and a hammock between two trees, star trail.		A bird known for its distinctive blue and orange plumage. The kingfisher is perched on a branch, its body angled slightly to the left as if poised to take flight at any moment.

**Fig. 3: Samples in our proposed High-Quality Evaluation Dataset.** The evaluation dataset presented in this paper contains samples of superior visual quality compared to those in COCO-30K.

**Table 3: Comparisons on FID and Clip-Score with Open-sourced T2I Models.** PIXART- $\Sigma$  demonstrates enhanced performance in terms of FID and Clip-Score on the curated 30K High-Quality Evaluation Dataset.

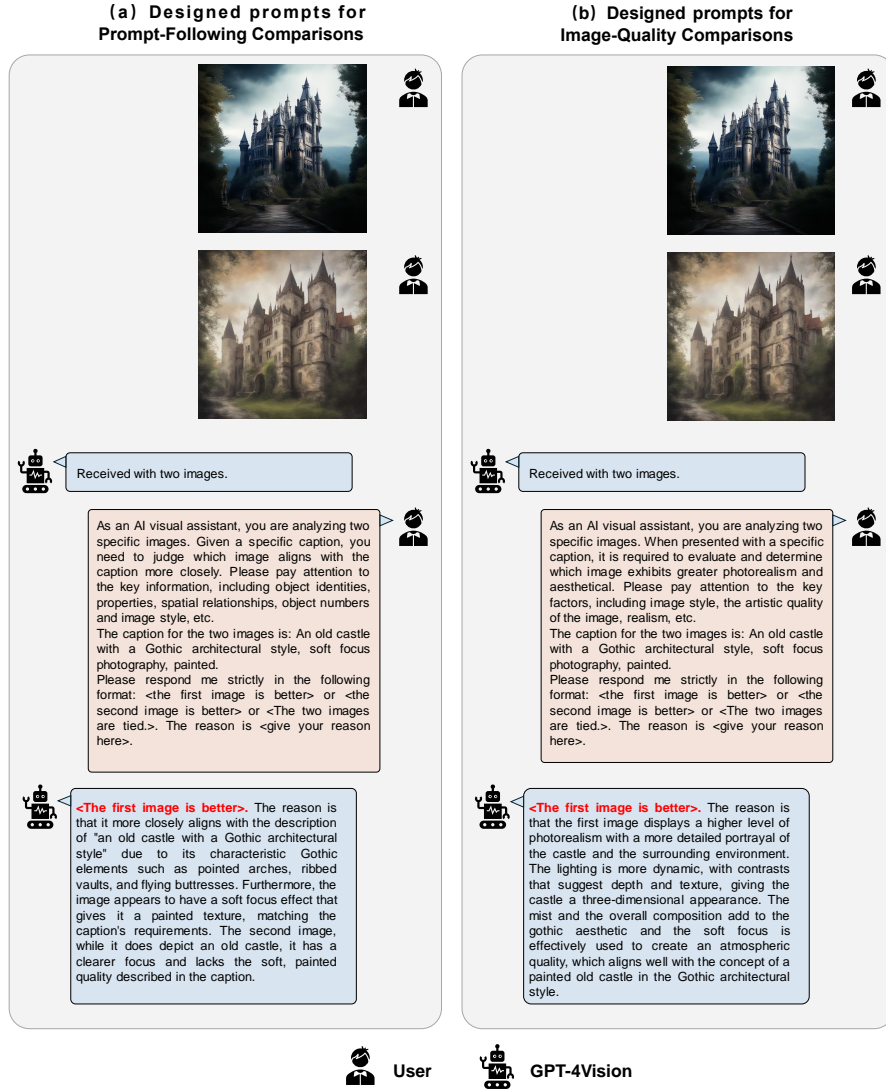
Models	#Params (B)	FID ↓	CLIP-Score ↑
Stable 1.5	0.9	17.03	0.2748
Stable Turbo	3.1	10.91	0.2804
Stable XL	2.6	7.38	0.2913
Stable Cascade	5.1	9.96	0.2839
Playground-V2.0	2.6	8.68	0.2885
Playground-V2.5	2.6	7.64	0.2871
PIXART- $\alpha$	0.6	8.65	0.2787
PIXART- $\Sigma$	0.6	8.23	0.2797

weak-to-strong fine-tuning process significantly improves the model’s ability to generate high-quality images and to align more closely with the given instructions, as PIXART- $\Sigma$  shows a lower FID (8.65 *v.s.* 8.23) and higher Clip-Scores (0.2787 *v.s.* 0.2797) compared to PIXART- $\alpha$ . Besides, compared to other models, the PIXART- $\Sigma$  still shows comparable or even better performance with relatively small network parameters (0.6B).

### 1.6 Designed Prompts for AI preference Study

To study AI preferences among various text-to-image (T2I) generators, we employ the advanced multi-modality model, GPT-4V, as an automated evaluator. The instructions given to GPT-4V for comparing the generated images based on quality and adherence to prompts are illustrated in Fig. 4. GPT-4V demonstrates the ability to provide logical assessments and reasons that coincide with human preferences.





**Fig. 4: Illustration of the designed prompts for GPT-4Vision.** We developed various prompts for GPT-4V to assess the image quality and instruction-following ability of different T2I generators, respectively.

### 1.7 Pseudo-code for KV Token Compression

We present the PyTorch-style pseudo-code for the KV Token Compression algorithm described in Sec.3.2 in the main paper. The implementation is straightforward.

**Algorithm 1** KV Token Compression.

---

```

import torch
import torch.nn as nn
import xformer

class AttentionKVCompress(nn.Module):
    def __init__(self, dim, sampling='conv', sr_ratio=1, **kwargs):
        super().__init__()

        # Projection layers and non-relevant definitions are omitted.
        self.sampling=sampling # ['conv', 'ave', 'uniform']
        self.sr_ratio = sr_ratio
        if sr_ratio > 1 and sampling == "conv":
            # Avg Conv Init.
            self.sr = nn.Conv2d(dim, dim, groups=dim, kernel_size=sr_ratio, stride=sr_ratio)
            self.sr.weight.data.fill_(1/sr_ratio**2)
            self.sr.bias.data.zero_()
            self.norm = nn.LayerNorm(dim)

    def downsample_2d(self, tensor, H, W, scale_factor, sampling=None):
        B, N, C = tensor.shape
        tensor = tensor.reshape(B, H, W, C).permute(0, 3, 1, 2)

        new_H, new_W = int(H / scale_factor), int(W / scale_factor)
        new_N = new_H * new_W

        if sampling == "ave":
            tensor = F.interpolate(
                tensor, scale_factor=1 / scale_factor, mode='nearest'
            ).permute(0, 2, 3, 1)
        elif sampling == "uniform":
            tensor = tensor[:, :, ::scale_factor, ::scale_factor].permute(0, 2, 3, 1)
        elif sampling == "conv":
            tensor = self.sr(tensor).reshape(B, C, -1).permute(0, 2, 1)
            tensor = self.norm(tensor)

        return tensor.reshape(B, new_N, C).contiguous(), new_N

    def forward(self, x, mask=None, HW=None, block_id=None, ):
        B, N, C = x.shape
        new_N = N
        H, W = HW
        qkv = self.qkv(x).reshape(B, N, 3, C)
        q, k, v = qkv.unbind(2)
        if self.sr_ratio > 1:
            k, new_N = self.downsample_2d(k, H, W, self.sr_ratio, sampling=self.sampling)
            v, new_N = self.downsample_2d(v, H, W, self.sr_ratio, sampling=self.sampling)

        x = xformers.ops.memory_efficient_attention(q, k, v)

        x = x.view(B, N, C)
        x = self.proj(x)
        x = self.proj_drop(x)
        return x

```

---

**1.8** PIXART- $\Sigma$  vs. T2I products

We compare PIXART- $\Sigma$  with four other close-source T2I products in Fig. 5, and Fig. 6. Our model can produce high-quality, photo-realistic images with rich details and is comparable with these products.

### 1.9 More images generated by PIXART- $\Sigma$

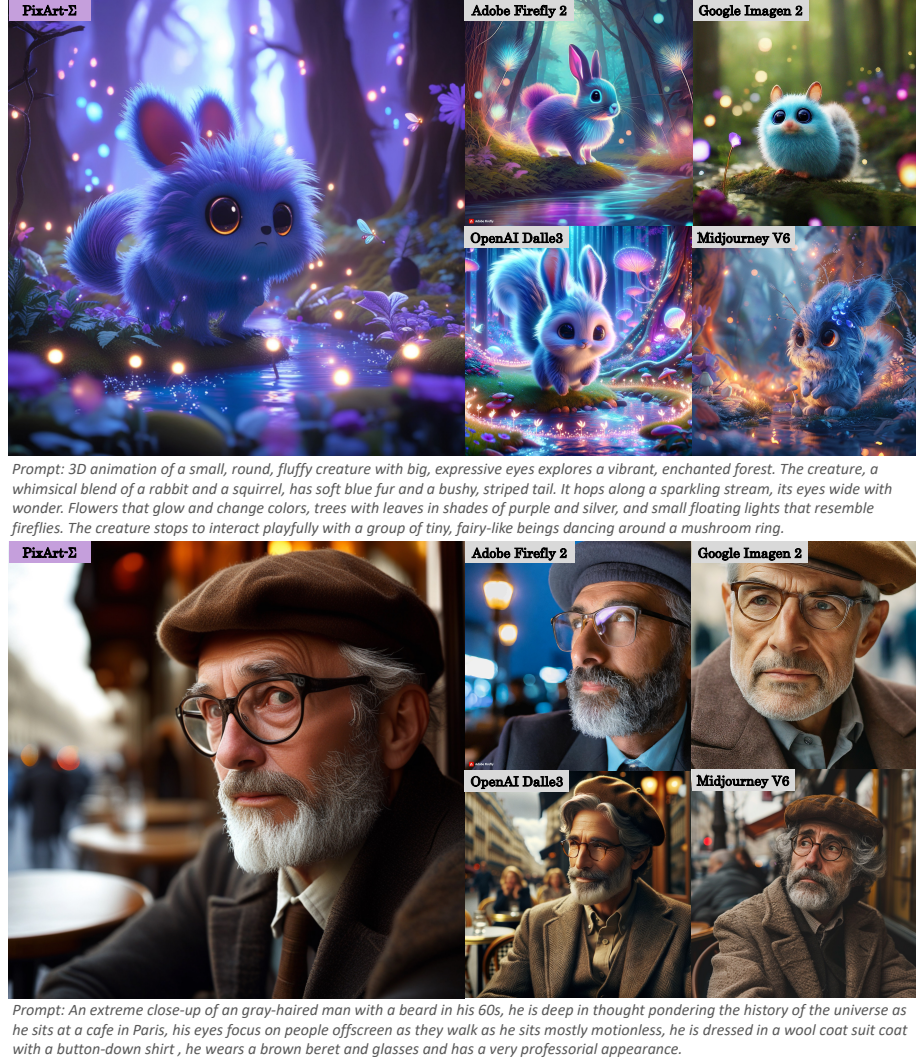
Fig. 7, 8 and 9 showcase additional visual outputs produced by PIXART- $\Sigma$ . The quality of these samples is remarkable, characterized by their high fidelity and accuracy in closely matching the provided textual prompts.

### 1.10 Limitation and Social Impact

**Limitation.** Our model still lacks the ability to generate some specific scenes and objects, especially text generation and hand generation. It is not perfect in the following aspects: it cannot fully align the complex prompts input by the user, face generation may have flaws and sensitive content may be generated. Subsequent research should focus on higher-quality data construction, scale model size, and improving model illusion and security issues through super alignment. **Negative social impact.** Text-to-image models may bring a negative social impact by generating images that present stereotypes or discriminate against certain groups. For instance, the images generated by a text-to-image model may depict unbalanced proportions of gender and inaccurate content for some uncommonly used concepts. Mitigating these issues requires careful data collection.

## References

1. Chen, J., Wu, Y., Luo, S., Xie, E., Paul, S., Luo, P., Zhao, H., Li, Z.: Pixart- $\delta$ : Fast and controllable image generation with latent consistency models. In: arXiv (2024) 1
2. Li, D., Kamko, A., Sabet, A., Akhgari, E., Xu, L., Doshi, S.: Playground v2, <https://huggingface.co/playgroundai/playground-v2-1024px-aesthetic> 3
3. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 3
4. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. CVPR (2024) 1



**Fig. 5: Compare PIXART- $\Sigma$  and four other T2I products:** Firefly 2, Imagen 2, Dalle 3, and Midjourney 6. Images generated by PIXART- $\Sigma$  are very competitive with these commercial products.





Prompt: A cute orange kitten sliding down an aqua slide. happy excited. 16mm lens in front. we see his excitement and scared in the eye. vibrant colors. water splashing on the lens

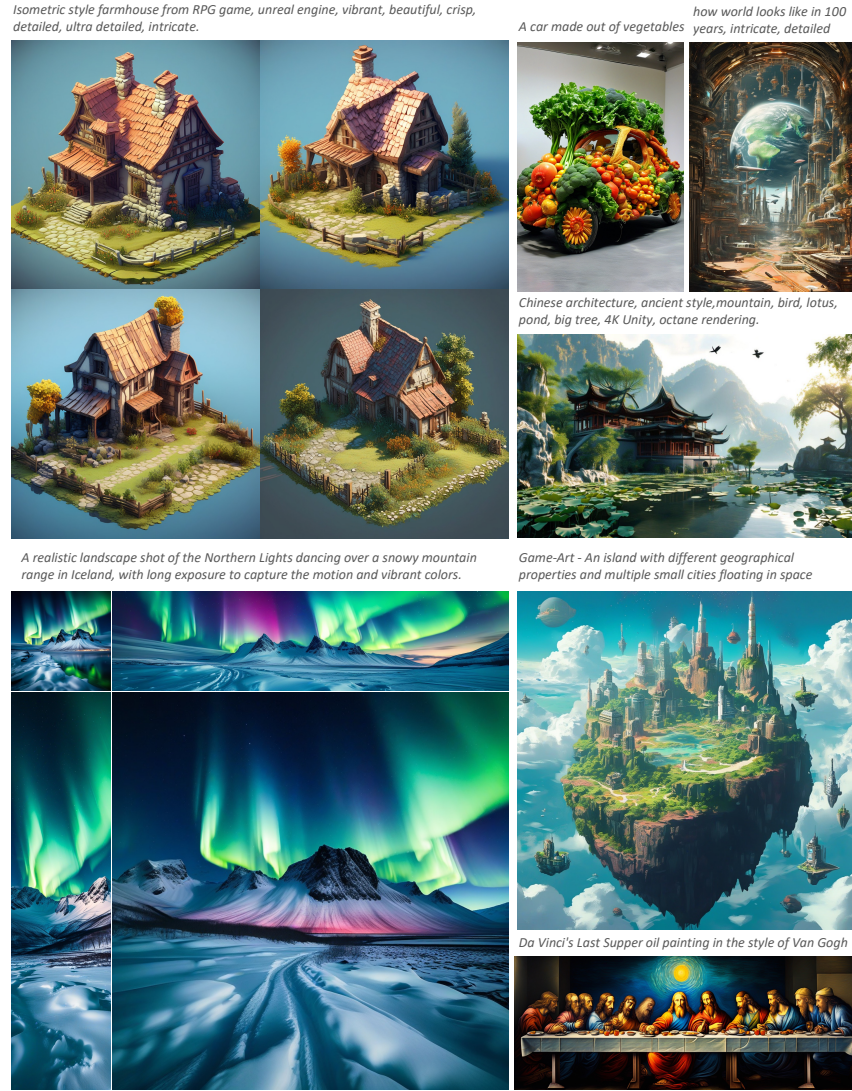


Prompt: a Chinese model is sitting on a train, magazine cover, clothes made of plastic, photorealistic, futuristic style, gray and green light, movie lighting, 32K HD



Prompt: several brightly colored rocks on a colorful beach, in the style of luminous spheres, 3840x2160, emek golan, translucent color, 32k uhd, toyn, captivating

**Fig. 6: Compare PIXART- $\Sigma$  and four other T2I products: Firefly 2, Imagen 2, Dalle 3, and Midjourney 6.** Images generated by PIXART- $\Sigma$  are very competitive with these commercial products.



**Fig. 7: Illustrations of High-quality images generated by PIXART- $\Sigma$ .** PIXART- $\Sigma$  can generate high-quality images with fine-grained details, and diverse images with different aspect ratios.



*Prompt: full body shot, a French woman, Photography, French Streets background, backlighting, rim light, Fujifilm*



**Fig. 8: High-resolution (4K) images generated by PIXART- $\Sigma$ .** PIXART- $\Sigma$  can directly generate high-quality 4K HD ( $3840 \times 2560$ ) images while preserving fine-grained details.

*Prompt: A deep forest clearing with a mirrored pond reflecting a galaxy-filled night sky*



**Fig. 9: High-resolution (4K) images generated by PixArt-Σ.** PixArt-Σ can directly generate high-quality 4K HD ( $3840 \times 2560$ ) images while preserving fine-grained details.