

## Abstract

In this paper, we introduce PixArt- $\Sigma$ , a Diffusion Transformer model (DiT) capable of directly generating images at 4K resolution. PixArt- $\Sigma$  represents a significant advancement over its predecessor, PixArt- $\alpha$ , offering images of markedly higher fidelity and improved alignment with text prompts. A key feature of PixArt- $\Sigma$  is its training efficiency. Leveraging the foundational pre-training of PixArt- $\alpha$ , it evolves from the ‘weaker’ baseline to a ‘stronger’ model via incorporating higher quality data, a process we term “weak-to-strong training”. The advancements in PixArt- $\Sigma$  are twofold: (1) High-Quality Training Data: PixArt- $\Sigma$  incorporates superior-quality image data, paired with more precise and detailed image captions. (2) Efficient Token Compression: we propose a novel attention module within the DiT framework that compresses both keys and values, significantly improving efficiency and facilitating ultra-high-resolution image generation. Thanks to these improvements, PixArt- $\Sigma$  achieves superior image quality and user prompt adherence capabilities with significantly smaller model size (0.6B parameters) than existing text-to-image diffusion models, such as SDXL (2.6B parameters) and SD Cascade (5.1B parameters). Moreover, PixArt- $\Sigma$ 's capability to generate 4K images supports the creation of high-resolution posters and wallpapers, efficiently bolstering the production of high-quality visual content in industries such as film and gaming.

**(本文贡献):**在本文中,我们介绍了 PixArt- $\Sigma$ , 一种扩散变换器模型 (DiT), 能够直接生成 4K 分辨率的图像。**(改进效果概述):**PixArt- $\Sigma$  代表了相比其前身 PixArt- $\alpha$  的一大进步, 提供了 **明显更高保真度的图像** 和与 **文本提示更好的对齐**。PixArt- $\Sigma$  的一个关键特性是其 **训练效率**。**(原理概述):**通过利用 PixArt- $\alpha$  的基础预训练, 它通过整合更高质量的数据, 从一个“较弱”的基线进化到一个“较强”的模型, 这一过程我们称之为“弱到强训练”。**(具体改进点):**PixArt- $\Sigma$  的进步主要体现在两个方面: (1) **高质量训练数据**: PixArt- $\Sigma$  整合了 **更高质量的图像数据**, 配对了 **更精确和详细的图像描述**。(2) **高效的 Token 压缩**: 我们在 **DiT 框架**内提出了一个 **新的注意力模块**, 它压缩了键和值, **显著提高了效率**并促进了超高分辨率图像的生成。**(改进之后的优势):**得益于这些改进, PixArt- $\Sigma$  以明显 **更小的模型大小 (0.6B 参数)** 达到了优于现有文本到图像扩散模型的图像质量和用户提示遵循能力, 如 SDXL (2.6B 参数) 和 SD Cascade (5.1B 参数)。此外, PixArt- $\Sigma$  生成 4K 图像的能力支持了高分辨率海报和壁纸的创作, 有效地加强了电影和游戏等行业中高质量视觉内容的生产。

Keywords: T2I Synthesis, Diffusion Transformer, Efficient Model

## 1 Introduction

**(研究背景):**近期, 高质量的文本到图像 (T2I) 模型的出现对人工智能生成内容 (AIGC) 社区产生了深远的影响。这包括专有模型如 DALL·E 3、Midjourney, 以及开源模型如 Stable Diffusion 和 PixArt- $\alpha$ 。**(先前研究的局限性):**然而, 开发一个顶级 T2I 模型需要 **大量资源**: 例如, 从头开始训练 SD1.5 大约需要 6000 个 A100 GPU 日, 这为拥有限制资源的个人研究者设置了一个重大障碍, 并阻碍了 AIGC 社区内的创新。随着时间的推移, AIGC 社区将获得持续更新的更高质量数据集和更先进的算法。**(科学问题):**一个关键的问题是: **我们如何在有限资源的约束下, 高效地将这些新元素整合到现有模型中, 实现一个更强大的版本?**

**(目的):**为了探讨这个问题, **(本文研究内容):**我们的研究聚焦于提升 PixArt- $\alpha$  的 **高效** 文本到图像 (T2I) 训练方法。PixArt- $\alpha$  代表了在 **扩散变换器 (DiT)** **框架**内的一次早期尝试, 这一模型结构具有显著的潜力, 如 Sora 和 Stable Diffusion 3 的研究所证实。为了最大化这一潜力, 我们在 PixArt- $\alpha$  的预训练基础上进行构建, 整合先进元素以促进其持续改进, 从而形成一个更强大的模型, PixArt- $\Sigma$ 。我们将这一从相对较弱的基线进化到更强大模型通过高效训练的过程称为“**弱到强训练**”。具体来说, 为了实现“弱到强训练”, 我们引入了以下改进:

**更高质量的训练数据**: 我们收集了一份比 PixArt- $\alpha$  中使用的数据更高质量的数据集, 重点关注两个关键方面: (i) **高质量图像**: 该数据集包含 33M 高分辨率图像, 均来源于互联网, 所有图像分辨率超过 1K, 其中包括 2.3M 分辨率约 4K 的图像。这些图像主要以其高审美特质为特点, 并涵盖了广泛的艺术风格。(ii) **密集准确的描述**: **为了为上述图像提供更精确和详细的描述**, 我们用更强大的 **图像描述生成器 Share-Captioner** 替换了 PixArt- $\alpha$  中使用的 LLaVA。此外, **(目的):**为了提高模型在文本和视觉概念之间的对齐能力, 我们将文本编码器 (即 Flan-T5) 的 token 长度扩展到大约 300 个词。我们观察到**(改进效果):**这些改进有效消除了模型的幻觉倾向, 导致更高质量的文本-图像对齐。

**(目的):**为了增强 PixArt- $\alpha$ ，我们将其生成分辨率从 1K 扩展到 4K。**(面临的问题):**在超高分辨率（例如，2K/4K）生成图像会显著增加 token 的数量，从而导致**计算需求大幅上升**。**(目的):**为了应对这一挑战，**(本文方法):**我们引入了一个**自注意力模块**，该模块配备了专为**扩散变换器（DiT）**框架量身定做的键（Key）和值（Value）token 压缩。具体来说，我们使用**步长为 2 的组卷积**来**局部聚合键和值**。此外，我们采用了一种专门的**权重初始化方案**，**允许从未使用 kv 压缩的预训练模型平滑过渡**。**(改进效果):**这种设计有效地减少了**高分辨率图像生成的训练和推理时间约 34%**。

**弱到强训练策略：**我们提出了几种微调技术，以便高效地从一个弱模型迅速适应到一个强模型。这包括**（1）**替换为更强大的变分自编码器（VAE），**（2）**从低分辨率到高分辨率的缩放，以及**（3）**从没有键值（KV）压缩的模型演进到具有 KV 压缩的模型。这些成果确认了“弱到强训练”方法的有效性和实用性。

**(最终效果):**通过提出的改进，PixArt- $\Sigma$ 在**最小的训练成本和模型参数**下实现了高质量 4K 分辨率图像生成。具体来说，从一个预训练模型进行微调，我们仅使用了 PixArt- $\alpha$  所需**GPU 日的 9%**就实现了一个强大的 1K 高分辨率图像生成模型，这在我们替换了新的训练数据和更强大的 VAE 的情况下尤其令人印象深刻。此外，我们仅使用了**0.6B** 参数，而 SDXL 和 SD Cascade 分别使用了**2.6B** 和**5.1B** 参数。由 PixArt- $\Sigma$  生成的图像具有与当前**顶级 T2I 产品**（如 DALL·E 3 和 MJV6）**相媲美的美学质量**（如图 4 所示）。此外，PixArt- $\Sigma$  还展示了与文本提示进行**精细对齐**的异常能力（如图 2 和 3 所示）。





图 1: 由 PixArt- $\Sigma$  生成的图像。该模型能输出具有照片级真实感、高审美价值、极端纵横比、多样风格的图像, 并且能够遵循用户指令。

A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.



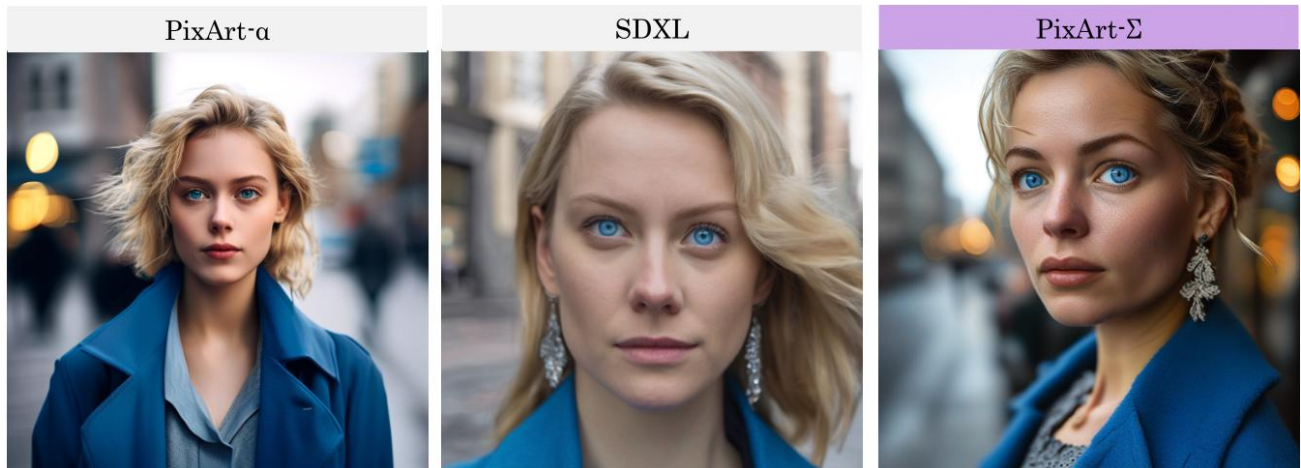
图 2: 应对复杂密集指令的 4K 图像生成。PixArt- $\Sigma$  能够直接生成 4K 分辨率图像, 无需后期处理, 并且能够准确响应给定的提示。

## 2 Related Work

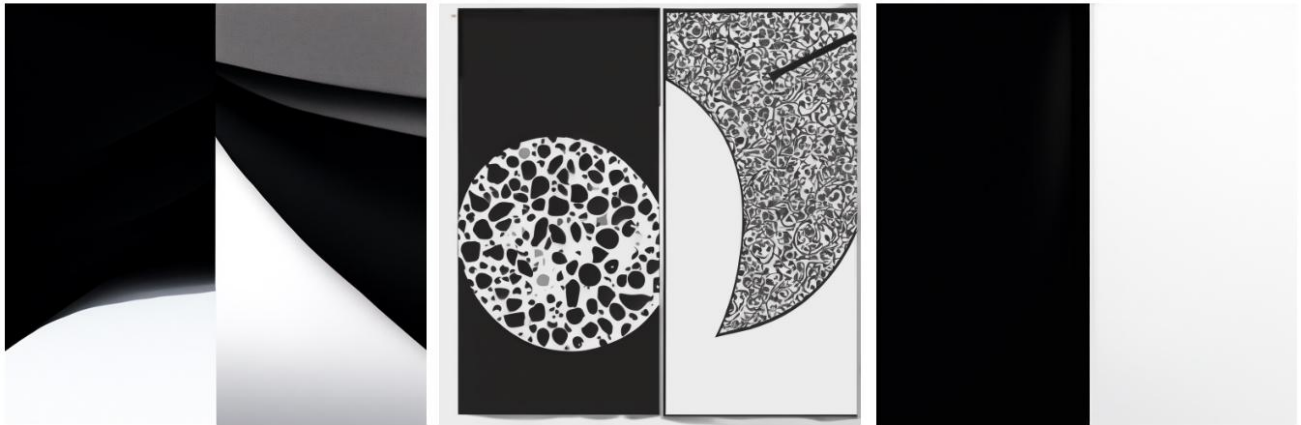
**扩散变换器。**Transformer 架构在各个领域都取得了显著的成功, 如语言建模、计算机视觉和其他领域。在扩散模型领域内, DiT 和 UViT 开创性地使用了 Transformer 架构。后续的工作, 包括 Diffit、SiT 和 FiT, 都在 DiT 的架构基础上进行了改进, 而另外一些工作则通过掩码建模技术提高了训练效率。在文本到图像 (T2I) 合成方面, PixArt- $\alpha$  探索了高效的 T2I 训练方案, 实现了第一个能够生成 1024px 高质量图像的基于 Transformer 的 T2I 模型。强大的视频生成模型 Sora 的出现进一步突显了扩散变换器的潜力。(本文贡献): 在这项工作中, 我们首次探索使用 Transformer 架构直接生成 4K 超高分辨率图像, 解决了涉及长序列 token 的计算复杂性挑战。

**高分辨率图像生成显著提升了视觉质量, 在电影和游戏等多个行业中十分重要。**然而, 提高图像分辨率由于计算需求的大幅增加而带来挑战。在这个方向上, 已经探索了许多方法。例如, Imagen、GigaGAN 和 Stable Diffusion 引入了额外的超分辨率网络, 而 Stable Cascade 使用多个扩散网络逐步提高分辨率。然而, (局限性): 这些组合模型解决方案可能会引入累积误差。另一方面, 如 SDXL、DALL·E 2、Playground 和 PixArt- $\alpha$  等工作尝试直接使用扩散模型生成高分辨率图像。尽管如此, (局限性): 这些努力由于计算复杂性而被限制在生成最高 1024px 分辨率的图像。(本文贡献): 在本文中, 我们将这一界限推向了 4K 分辨率, 显著提升了生成内容的视觉质量。

**高效的 Transformer 架构。**Transformer 中的自注意力机制随着 token 数量的增加呈二次方计算复杂度增长, 这限制了 token 数量的扩展。许多工作在这一领域寻求改进: (1) 稀疏注意力, 通过选择性地处理一部分 token 来减少整体的计算负担。例如, PVT v2 使用卷积核压缩 key 和 value 的空间, 从而降低了计算注意力的复杂性。(2) 局部注意力, 关注近邻区域内的信息; 值得注意的是, Swin Transformer 利用基于窗口的注意力限制了在指定窗口大小内的计算。(3) 低秩/线性注意力。Linformer 通过低秩近似减少了自注意力机制的计算复杂性。在本文中, 受到 PVT v2 的启发, 我们采用了基于键/值压缩的自注意力机制, 以减轻处理 4K 图像的高复杂性。



Prompt: A close-up photo of a person. The subject is a woman. She wore a **blue coat** with a **gray dress underneath**. She has **blue eyes** and **blond hair**, and wears a pair of **earrings**. Behind are blurred city buildings and streets.



Prompt: half a **solid black** background and half a **solid white** background



Prompt: **Pixel art style** of a snowboarder **in mid-air** performs a trick **on a black rail**, wearing a **blue sweatshirt** and **black pants**, with **arms outstretched**. The serene snowy landscape background, dotted with trees, complements the scene. The low-angle perspective emphasizes the trick's height and skill.

**图 3:** PixArt- $\Sigma$  与开源模型（例如，PixArt- $\alpha$  和 SDXL）的比较：与 PixArt- $\alpha$  相比，PixArt- $\Sigma$  提高了肖像的真实感和语义分析的能力。与 SDXL 相比，我们的方法更能够遵循用户指令。关键词以蓝色高亮显示。

### 3 Framework



### 3.1 Data Analysis

**更高的审美和更高的分辨率。** **(目的):** 为了提升我们数据集的审美质量, **(本文方法):** 我们将内部数据从 1400 万 **扩展** 到 3300 万。为清晰起见, 我们分别将这两个数据集命名为内部- $\alpha$  和内部- $\Sigma$ 。 **(局限性):** 需要注意的是, 与目前可用的开源模型 (如 SD v1.5 使用的 20 亿数据) 相比, 这种 **扩展仍显不足**。 **(本文结论):** 我们证明了即便是数据量有限, 通过有效的训练策略依然可以获得强大的文本到图像 (T2I) 模型。

内部- $\Sigma$  中的图像分辨率超过 1K。 **(目的):** 为了实现 4K 分辨率生成, 我们额外收集了 800 万张真实的 4K 分辨率摄影图像。 **(目的):** 为了确保审美质量, 我们采用了一个 **审美评分模型** (AES) [1] 来过滤这些 4K 图像。 **(效果):** 这一过程产生了一个 **高度精炼** 的 200 万 **超高分辨率且高质量图像** 的数据集。



Prompt: a small cactus with a happy face in the Sahara desert



Prompt: a cherry pick scientist

**图 4:** 将 PixArt- $\Sigma$  与其他四款文本到图像 (T2I) 产品进行比较: Firefly 2、Imagen 2、Dalle 3 和 Midjourney 6。由 PixArt- $\Sigma$  生成的图像与这些商业产品具有很强的竞争力。



	<b>LLaVA Hallucinations</b> The image features a large, ornate church with a tall, <b>pointed roof</b> and a large <b>stained-glass window</b> . The church is situated in a city, and it has a white and gray color scheme. The style of the church is Gothic, which is evident from the pointed roof and the intricate details of the <b>stained-glass window</b> . The presence of statues and a <b>clock</b> on the building further emphasizes the grand and historical nature of the structure.
	<b>Share-Captioner Correctness</b> The image captures the grandeur of a cathedral, painted in a pristine white, stands majestically against the backdrop of a <b>clear blue sky</b> . The entrance to the cathedral is adorned with a <b>large rose window</b> . Flanking the entrance are <b>statues of saints</b> , standing as silent guardians of the sacred space within. The perspective of the photo, <b>taken from a low angle</b> . The <b>lush green trees</b> in the background adding serenity to the scene.
	<b>LLaVA Hallucinations</b> The image features a <b>woman and a man</b> sitting on a brick walkway near a body of water, which could be a river or a lake. They are both wearing head coverings, and the <b>woman is holding a handbag</b> . The scene is set during the day, with the <b>sun shining brightly</b> , creating a warm and inviting atmosphere. The style of the image is a <b>black and white photo</b> , which adds a <b>timeless and classic</b> feel to the scene.
	<b>Share-Captioner Correctness</b> The image captures a serene scene at a <b>harbor</b> . <b>Two individuals</b> are seated on a bench, their backs to the camera, engrossed in the view of the water. The water, a deep shade of blue, is dotted with <b>boats of various sizes and colors</b> , including a <b>white boat with a green stripe</b> and a <b>red boat</b> . The sky above is a light blue.

图 5: 幻觉对比图示: 通过红色表示幻觉、绿色表示正确性, 对 LLaVA 与 Share-Captioner 之间发生幻觉的差异进行对比。

(发现): 有趣的是, 我们观察到随着图像分辨率的提高, 模型的保真度 (Fréchet Inception Distance (FID) [16]) 和语义对齐 (CLIP 得分) 也有所改善, 这强调了生成高分辨率图像能力的重要性。

**更好的文本-图像对齐。**近期的作品, 如 PixArt- $\alpha$  [4]和 DALL-E 3 [30], 强调了文本-图像描述对齐的重要性。加强这种对齐对于提升模型能力至关重要。为了进一步精炼我们收集的“原始”描述, 我们专注于提高描述的长度和准确性。值得注意的是, 我们的描述 (内部- $\Sigma$ ) 在以下几个方面相比于 PixArt- $\alpha$  (内部- $\alpha$ ) 使用的描述显示出几个优势:

- 1. 增强的描述准确性:** 如图 5 所示, PixArt- $\alpha$  中使用的 LLaVa 存在一定的幻觉问题。我们利用一个更强大的视觉语言模型, 即 Share-Captioner [5], 来生成详细且正确的描述, 增强了收集到的原始提示。
- 2. 增加的描述长度:** 如表 1 和图 6 所展示的, 平均描述长度显著增加到 180 个词, 极大地增强了描述的表达能力。此外, 我们将文本编码器的令牌处理长度从 120 个令牌 (如同内部- $\alpha$  中的设置) 扩展到 300 个令牌。我们的模型训练了一种长 (Share-Captioner) 与短 (原始) 描述的混合, 比例分别为 60% 和 40%。(效果): 这种方法增加了文本描述的多样性, 并减轻了仅依赖生成描述可能产生的潜在偏见。

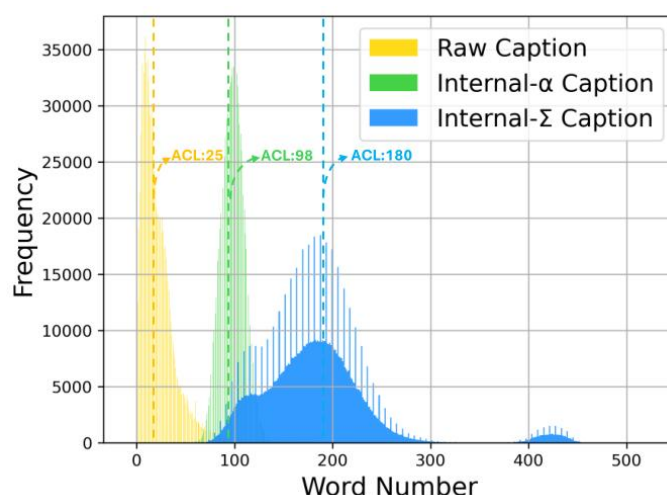


图 6: 描述长度的直方图可视化。我们随机从原始描述、内部- $\alpha$  和内部- $\Sigma$  中选取 100 万个描述来绘制相应的直方图。ACL 代表平均描述长度。

表 1 展示了内部- $\alpha$  和- $\Sigma$  的总结, 我们通过包括名词多样性、总名词数量、平均描述长度和每张图像平均名词数量等多个指标 (目的): 来评估数据集



的多样性。

**高质量评估数据集。**大多数最先进的文本到图像（T2I）模型选择了 **MSCOCO** [20]作为评估集来评估 **FID** 和 **CLIP 得分**。**(局限性):**然而，我们观察到在 MSCOCO 数据集上进行的评估 **可能不足以充分反映一个模型在审美和文本-图像对齐方面的能力**。**(本文贡献):**因此，我们提出了一个由 30,000 个高质量、审美上令人满意的文本-图像对组成的精选集，以便于进行评估。所选样本在附录中呈现。这个数据集旨在提供一个更全面的模型性能评估，特别是在捕捉审美吸引力的复杂性和文本描述与视觉内容之间对齐的保真度方面。除非另有说明，论文中的评估实验都是在收集的高质量评估数据集上进行的。

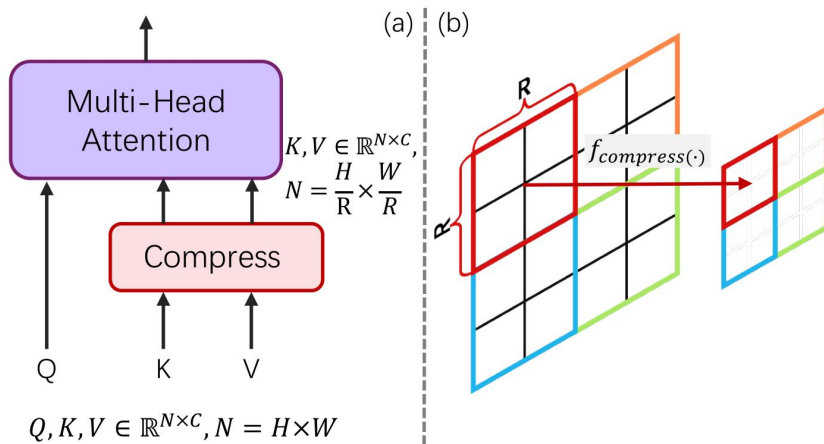
Dataset	Volume	Caption	VN/DN	Total Noun	ACL	Average
Internal- $\alpha$	14M	Raw	187K/931K	175M	25	11.7/Img
Internal- $\alpha$	14M	LLaVA	28K/215K	536M	98	29.3/Img
Internal- $\alpha$	14M	Share-Captioner	51K/420K	815M	184	54.4/Img
Internal- $\Sigma$	33M	Raw	294K/1512K	485M	35	14.4/Img
Internal- $\Sigma$	33M	Share-Captioner	77K/714K	1804M	180	53.6/Img
4K- $\Sigma$	2.3M	Share-Captioner	24K/96K	115M	163	49.5/Img

**表 1:** 不同数据集名词概念的统计数据。VN: 有效的不同名词（出现超过 10 次）；DN: 总的不同名词；平均: 每张图像的平均名词数量；ACL: 平均描述长度。

### 3.2 Efficient DiT Design

高效的 DiT（Diffusion Transformer）网络是至关重要的，因为当生成超高分辨率图像时，**计算需求显著增加**。**注意力机制**在扩散变换器的效率中扮演了关键角色，**(局限性):**然而其二次方的计算需求显著限制了模型的可扩展性，特别是在更高的分辨率下，例如 2K 和 4K。**(本文方法):**受到 PVT v2 [45] 的启发，我们在原始的 PixArt- $\alpha$  框架内**整合了 KV 压缩**，以应对计算挑战。**(效果):**这种设计仅增加了总参数的 0.018%，但通过令牌压缩实现了有效的**计算成本降低**，同时仍然保留了空间和语义信息。

**键值 (KV) 令牌压缩。**我们的动机来源于一个有趣的观察：**(出发点):**直接将键值 (KV) 令牌压缩应用于预训练的 PixArt- $\alpha$  仍然可以生成合理的图像。**(结论):**这表明**特征中存在冗余**。考虑到相邻  $R \times R$  块内高度相似性，我们假设一个窗口内的特征语义是冗余的，并且可以合理地压缩。**(本文方法):**我们提出了 **KV 令牌压缩**，记为  $f_c(\cdot)$ ，通过一个压缩操作符来压缩  $R \times R$  窗口内的令牌特征，如图 7 所示。



**图 7:** KV 令牌压缩设计。我们在空间上合并 KV 令牌以降低计算复杂性。

进一步地, (目的):为了减轻 kv 压缩在自注意力计算中可能造成的信息损失, (本文方法):我们选择保留所有查询 (Q) 的令牌。(效果):这一策略使我们能够有效地利用 kv 压缩, 同时减少丢失关键信息的风险。(最终 kv 压缩效果):通过采用 kv 压缩, 我们提高了注意力计算的效率, 并将计算复杂性从  $O(N^2)$ 降低到  $O(N^2/R^2)$ , 从而使直接生成高分辨率图像的计算成本变得可管理。


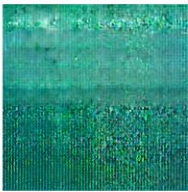


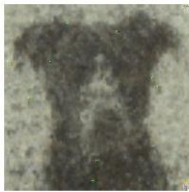

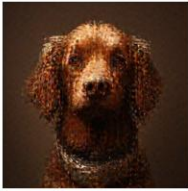

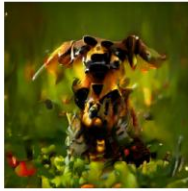

Strategies	VAE Change	512 →1024 RA	512 →1024 RA + PE Interp.	RA + PE Adjust + KV Compress	RA + PE Interp. + KV Compress + Conv Avg Init
0 step ↓ Fast Adapt					
	 2K step	 100 step	 100 step	 100 step	 100 step
(a)		(b)		(c)	

图 8: 此示意图展示了我们的训练策略设计如何在过渡到 VAE、调整到更高分辨率和 kv 压缩过程中加速模型的收敛, 促进从弱到强的快速学习。

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{Q \cdot f_c(K)^T}{\sqrt{d_k}} \right) f_c(V) \tag{1}$$

我们使用具有特定初始化的卷积操作符 “Conv2×2” 来压缩深层。在第 5 节中讨论了其他设计变体的详细实验。(本文创新):具体来说, 我们设计了一种专门的卷积核初始化方法 “Conv Avg Init”, 该方法利用群卷积并将权重  $w$  初始化为  $\frac{1}{R^2}$ , 相当于一个平均操作符。(效果):这种初始化策略最初可以产生粗糙的结果, 加速微调过程, 同时仅引入了 0.018% 的额外参数。

3.3 Weak-to-Strong Training Strategy

我们提出了几种高效的训练策略来增强从 “弱” 模型到 “强” 模型的过渡。这些策略包括 VAE 快速适应、高分辨率微调和 kv 令牌压缩。

适应新的 VAEs。随着 VAEs 的不断发展, 从头开始训练 T2I 模型是资源密集型的。(本文方法):我们用 SDXL 的 VAE 替换了 PixArt-α 的 VAE, 并继续对扩散模型进行微调。(发现):我们观察到一个快速收敛现象, 即微调在 2000 个训练步骤时迅速收敛, 如图 8 (a) 所示。当处理 VAE 模型转移时, 微调更为高效, 并且避免了从头开始训练的必要性。

Resolution	Iterations	FID ↓	CLIP ↑
256	20K	16.56	0.270
256 → 512	1K	9.75	0.272
256 → 512	100K	8.91	0.276

表 2: 我们从低分辨率模型微调一个高分辨率模型, 并观察到即使是相对较短的微调时长, 如 1000 步, 仍然可以获得高质量的结果。

适应更高分辨率。(发现):当我们从低分辨率 (LR) 模型微调到高分辨率 (HR) 模型时, 我们观察到性能下降, 如图 8 (b) 所示, (解释):我们将其归因于不同分辨率间位置嵌入 (PE) 的差异。(目的):为了解决这个问题, (本文方法):我们使用了 “PE 插值” 技巧 [4,48]: (效果):通过插值 LR 模型的 PE 来初始化 HR 模型的 PE, 显著增强了 HR 模型的初始状态并加速了微调过程。即使只在 100 个训练迭代内, 我们也能获得视觉上令人满意的图像。(定



**量评估结果):**此外,我们如表 2 所示定量评估模型性能的变化。微调在 1000 步时迅速收敛,进一步训练稍微提高了性能。**(结论):**这说明使用“PE 插值”技巧能够实现更高分辨率生成的快速收敛,避免了从头开始训练生成更高分辨率图像的需要。

**适应 KV 压缩模型。**我们可以在从未使用 KV 压缩的低分辨率预训练模型进行微调时直接使用 KV 压缩。如图 8 (c) 所示,通过我们的“Conv Avg Init.”策略,**(效果):**PixArt- $\Sigma$  从一个更好的初始状态开始,使得**收敛变得更容易且更快**。值得注意的是,即使在 100 个训练步骤内, PixArt- $\Sigma$  也能呈现满意的视觉结果。最终,通过在第 3.2 节中设计的 KV 压缩操作符和压缩层,我们可以减少大约 34% 的训练和推理时间。

## 4 Experiment

### 4.1 Implementation Details

**训练细节。**我们遵循 Imagen [39] 和 PixArt- $\alpha$  [4] 的做法,使用 T5 [9] 的编码器(即 Flan-T5-XXL)作为条件特征提取的文本编码器,并使用 PixArt- $\alpha$  [4] 作为我们的基础扩散模型。**(本文方法):**与大多数工作使用固定的 77 个文本令牌不同,我们将文本令牌的长度从 PixArt- $\alpha$  的 120 调整到 300,因为在内部- $\Sigma$  中策划的描述更加密集,**(目的):**以提供高度细致的细节。**(目的):**为了捕获输入图像的潜在特征,**(本文方法):**我们使用了来自 SDXL [35] 的预训练且冻结的 VAE。其他实现细节与 PixArt- $\alpha$  相同。模型在 PixArt- $\alpha$  的 256px 预训练检查点上进行调整,采用位置嵌入插值技巧[4]。我们的最终模型,包括 1K 分辨率的,是在 32 个 V100 GPU 上训练的。我们还使用了 16 个 A100 GPU 来训练 2K 和 4K 图像生成模型。更多信息,请参见附录。

注意,我们使用 CAME 优化器[26],权重衰减为 0,学习率恒定为 2e-5,而不是常规的 AdamW [23] 优化器。**(效果):**这有助于我们减少优化器状态的维度,从而降低 GPU 内存而不影响性能。

**评估指标。****(目的):**为了更好地展示美学和语义能力,**(本文方法):**我们收集了 30K 高质量的文本-图像对(如第 3.1 节所述)**(目的):**来基准测试最强大的 T2I 模型。**(本文方法):**我们主要通过人类和 AI 偏好研究来评估 PixArt- $\Sigma$ ,**(本文观点):**因为 FID [38] 指标可能不充分反映生成质量。然而,我们仍然在附录中提供了在收集的数据集上的 FID 结果。

### 4.2 Performance Comparisons

**图像质量评估。**我们对我们的方法进行了定性评估,与闭源的文本到图像(T2I)产品和开源模型进行了比较。如图 1 所示,**(本文模型效果):**我们的模型可以生成高质量、逼真的图像,这些图像在不同的纵横比和风格上都具有复杂的细节。**(结论):**这一能力突显了我们的方法在从文本描述生成视觉上引人入胜的内容方面的卓越性能。如图 3 所展示的,我们将 PixArt- $\Sigma$  与开源模型 SDXL [35] 和 PixArt- $\alpha$  [4] 进行了比较,**(本文模型效果):**我们的方法提高了肖像的真实感并增强了语义分析的能力。与 SDXL 相比,我们的方法在遵循用户指令方面显示出了更高的熟练度。

我们的方法不仅在开源模型上有优势,而且在当前的 T2I 闭源产品中也非常具有竞争力,如图 4 所描绘的。PixArt- $\Sigma$  产生的是逼真的结果,并且紧密遵循用户指令,这与当代的商业产品不相上下。

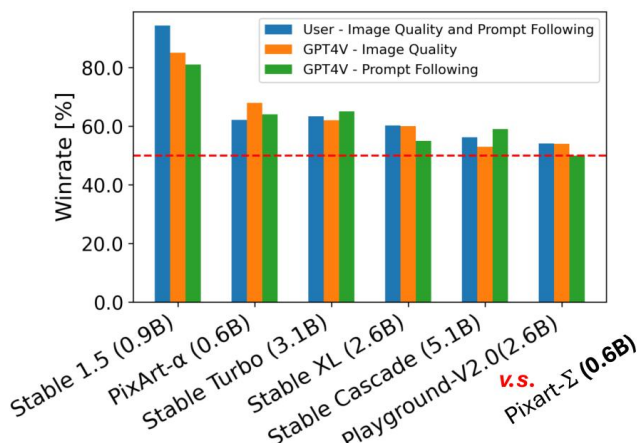


图 9: 人类(蓝色)/AI(橙色和绿色)偏好评估与当前开放 T2I 模型的比较。PixArt- $\Sigma$  在图像质量和遵循提示方面都优于当前最先进的 T2I 模型。

**高分辨率生成。(本文方法优势):**我们的方法能够**直接生成 4K 分辨率图像**，无需任何后期处理。此外，它在准确遵循用户提供的复杂、详细和长文本方面表现出色，如图 2 所示。因此，用户**不需要进行提示工程**就能获得满意的结果。我们的方法实现了直接 4K 图像生成。**(过去研究):**同时，研究[10, 15]引入了无需调整的后处理技术，旨在从低分辨率模型生成高分辨率（HR）图像，或使用超分辨率模型[49]产生 HR 图像。**(过去研究的局限性):**然而，他们对应的结果经常出现**人工痕迹**，主要有两个原因：**(1)** 由于**级联管道**可能产生累积错误。**(2)** 这些方法没有捕捉到 4K 图像的真实分布，也没有学习文本与 4K 图像之间的对齐。**(本文观点):**我们认为，我们的方法可能是生成高分辨率图像的一种更有希望的方式。我们的方法产生了更优越的结果，并且在补充材料中包含了更多的视觉比较。

**人类/AI (GPT4V) 偏好研究。**我们使用第 3.1 节提到的高质量评估数据集中随机收集的 300 个描述的子集，对经过良好训练的模型进行了人类和 AI 偏好研究。我们收集了包括 **PixArt- $\alpha$** 、**PixArt- $\Sigma$** 、**SD1.5** [38]、**Stable Turbo** [40]、**Stable XL** [35]、**Stable Cascade** [34]和 **Playground-V2.0** [19]在内的六个开源模型生成的图像。**(本文贡献):**我们开发了一个网站用于**人类偏好研究**，展示提示及其对应的图像。

这个网站被分发给训练有素的**评估员**，他们被要求评估图像，并根据质量和图像与文本提示的匹配程度进行排名。**(评估结果):**结果，如图 9 中的蓝色条形图所示，标志着对 PixArt- $\Sigma$ 相比其他六个 T2I 生成器的明显偏好。**(本文模型优势):**PixArt- $\Sigma$ 生成了高质量的图像，这些图像紧密遵循用户提示，与现有的 T2I 扩散模型（如 SDXL（2.6B 参数）和 SD Cascade（5.1B 参数））相比，使用了**更小的大小**（0.6B 参数）。

此外，在我们的 AI 偏好研究中，**(评估方式):**我们使用高级多模态模型 GPT-4 Vision [31]作为评估者。在每次试验中，我们为 GPT-4 Vision 提供两张图像：一张来自 PixArt- $\Sigma$ ，另一张来自竞争的 T2I 模型。我们设计了独特的提示，指导 GPT-4 Vision 根据图像质量和图文对齐程度进行投票。**(评估结果):**结果，由图 9 中的橙色和绿色条表示，展示了在人类和 AI 偏好研究中一致的结果。**(本文模型优势):**特别是，PixArt- $\Sigma$ 在效果上超过了基线模型 PixArt- $\alpha$ 。与如 Stable Cascaded 这样的当代高级模型相比，PixArt- $\Sigma$ 在**图像质量**和**指令遵循能力**方面展示了**有竞争力或更优的性能**。

## 5 Ablation Studies

我们对各种 **kV 压缩设计**的生成性能进行了**消融研究**。除非另有说明，实验都是在 512px 生成上进行的。每个消融实验的详细设置都包含在附录中。

### 5.1 Experimental settings

我们使用第 3.1 节中描述的测试集进行评估。我们采用 **FID** 来计算收集的数据和生成的数据之间的**分布差异**，用作比较指标。此外，我们利用 **CLIP-Score** 来评估提示与生成图像之间的**对齐情况**。

Layers	FID ↓ CLIP-Score ↑	
N/A	8.244	0.276
Shallow (1-14)	9.278	0.275
Middle (7-20)	9.063	0.276
Deep (14-27)	8.532	0.275

(a) Compression layers.

Operator	FID ↓ CLIP-Score ↑	
N/A	8.244	0.276
Token Discarding	8.918	0.275
Token Pooling	9.415	0.275
Conv2×2	8.505	0.274

(b) Compression operators.

Res.	Ratio	FID ↓	CLIP-Score ↑	Train Latency ↓
512	1	8.244	0.276	2.3
512	2	9.063	0.276	2.2 (-4%)
512	4	9.606	0.276	2.1 (-9%)
1024	1	5.685	0.277	27.5
1024	2	5.512	0.273	22.5 (-18%)
1024	4	5.644	0.276	20.0 (-27%)
1024	9	5.712	0.275	17.8 (-35%)

(c) Compression ratios on different resolutions.

Res.	Ratio	Train Latency ↓ (s/Iter@32BS)	Test Latency ↓ (s/Img)
2K	1	56	58
2K	4	37 (-34%)	38 (-34%)
4K	1	191	91
4K	4	125 (-35%)	60 (-34%)

(d) Speed of different resolutions.

**表 3:** 图像生成中的 KV-令牌压缩设置。本研究使用 FID、CMMD 和 CLIP-Score 指标来评估各种令牌压缩组件的影响，例如压缩比例、位置、操作符以及不同的分辨率。表 3c 中的速度计算是秒/迭代/384 批量大小。



5.2 Compression Designs

**压缩位置。**我们在 Transformer 结构的 **不同深度** 实现了 KV 压缩：浅层（1~ 14）、中间层（7~ 20）和深层（14~ 27）。(结果):如表 3a 所示，在 **深层** 使用 KV 压缩明显实现了 **更优** 的性能。(解释):我们推测这是因为浅层通常编码细节纹理内容，而 **深层抽象高级语义内容**。由于压缩倾向于 **影响图像质量而非语义信息**，压缩深层可以 **实现最小的信息损失**，使其成为 **加速训练** 而不损害生成质量的实用选择。

**压缩操作符。**我们探索了 **不同压缩操作符** 的影响。我们采用了三种技术，**随机丢弃**、**平均池化** 和 **参数化卷积**，将 2×2 令牌 **压缩** 成一个令牌。(结果):如表 3b 所示，“Conv 2×2”方法 **优于其他方法**，强调了 **使用可学习核心以更有效地减少冗余特征比简单丢弃方法的优势**。

**不同分辨率的压缩比。**我们调查了 **不同压缩比** 对不同分辨率的影响。如表 3c 所示，值得注意的是，(发现):我们发现 **令牌压缩不影响文本与生成图像之间的对齐（CLIP 得分）**，但 **影响不同分辨率下的图像质量（FID）**。(策略优势):尽管随着压缩比的增加图像质量略有下降，我们的策略带来了 **18% 至 35% 的训练加速**。(结论):**这表明我们提出的 KV 压缩既有效又高效，适用于实现高分辨率 T2I 生成。**

**不同分辨率下的速度比较。**我们进一步全面验证了训练和推理中的速度加速，见表 3d。我们的方法可以在 4K 生成中将训练和推理 **速度提高约 35%**。值得注意的是，(发现):我们观察到 **随着分辨率的提高，训练加速度也在增加**。例如，随着分辨率从 1K 增加到 4K，训练逐渐从 18% 加速到 35%。(结论):**这表明我们的方法随着分辨率的增加而有效，展示了其潜在适用性，甚至适用于更高分辨率图像生成任务。**

6 Conclusion

(本文贡献):在本文中，我们介绍了 PixArt-Σ，一个能够直接生成 4K 分辨率高质量图像的文本到图像（T2I）扩散模型。(模型构建概述):PixArt-Σ 基于 PixArt-α 的预训练基础，通过一种新颖的“从弱到强训练”方法实现了高效训练。(方法特点):这种方法的特点是结合了 **更高质量的数据和高效的令牌压缩集成**。(模型效果优势):PixArt-Σ 擅长生成 **高保真图像**，同时 **紧密遵循文本提示**，超越了其前身 PixArt-α 设定的高标准。(本文研究意义):我们相信，PixArt-Σ 中呈现的创新不仅将有助于推动人工智能生成内容（AIGC）社区的进步，还将为实体提供访问更高效、高质量生成模型的途径。

A Appendix

A.1 Training Details

在表 4 中，我们提供了 PixArt-Σ 每个训练阶段的详细信息，包括图像分辨率、训练样本总量、训练步数、批量大小、学习率和以 GPU 天计算的计算时间。利用内部-Σ 数据集并集成了 **更先进的 VAE**，我们提出的方法仅使用 5 个 V100 GPU 天就快速适应了新的 VAE。随后，我们仅使用 50 个 V100 GPU 天就实现了卓越的文本-图像对齐。

值得注意的是，应用 **KV 令牌压缩** 作为一个 **显著的效率提升器**，**大大减少了训练持续时间**。例如，从 512px 微调到 1024px 并实施 KV 压缩，所需的训练时间从 50 个 V100 GPU 天减少到仅 20 个 V100 GPU 天。同样，对于 2K 和 4K 的分辨率，训练时间分别从 10 个 A800 GPU 天减少到 7 个 A800 GPU 天，以及从 16 个 A800 GPU 天减少到 12 个 A800 GPU 天。这展示了 KV 令牌压缩在提高训练效率方面的有效性。

Stage	Image Resolution	#Images	Training Steps	Batch Size	Learning Rate	GPU days
VAE adaption	256×256	33M	8K	64×16	$2\times10^{-5}$	5 V100
Better Text-Image align	256×256	33M	80K	64×16	$2\times10^{-5}$	50 V100
Higher aesthetics	512×512	18M	10K	32×32	$2\times10^{-5}$	30 V100
Higher aesthetics	1024×1024	18M	5K	12×32	$1\times10^{-5}$	50 V100
KV token compression	1024×1024	18M	5K	12×16	$1\times10^{-5}$	20 V100
Higher aesthetics	2K×2K	300K	2K	4×8	$2\times10^{-5}$	10 A800
KV token compression	2K×2K	300K	2K	4×8	$2\times10^{-5}$	7 A800
Higher aesthetics	4K×4K	100K	1K	4×8	$2\times10^{-5}$	16 A800
KV token compression	4K×4K	100K	1K	4×8	$2\times10^{-5}$	12 A800

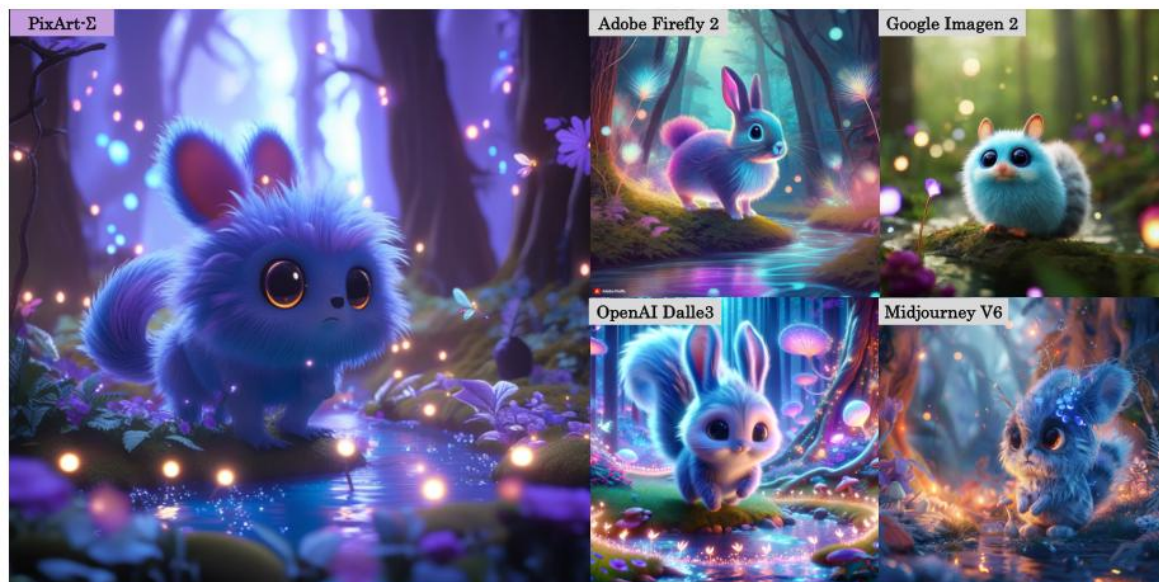
表 4：我们在论文中报告了 PixArt-Σ 每个训练阶段的详细信息。请注意，这里的内部-Σ 数据集包括 3300 万条内部数据。GPU 天数不包括 VAE 特征提取和 T5 文本特征提取的时间，因为我们提前离线准备了这两种特征，所以它们不是训练过程的一部分，也不会给它增加额外的时间。

## A.2 PixArt- $\Sigma$ vs. T2I products

我们在图 10 和图 11 中将 PixArt- $\Sigma$  与其他四个闭源文本到图像（T2I）产品进行了比较。我们的模型能够生成高质量、逼真的图像，拥有丰富的细节，并且可以与这些产品媲美。

## A.3 More images generated by PixArt- $\Sigma$

图 12、14 和 13 展示了由 PixArt- $\Sigma$  产生的额外视觉输出。这些样本的质量令人瞩目，其高保真度和在紧密匹配提供的文本提示方面的准确性特别突出。



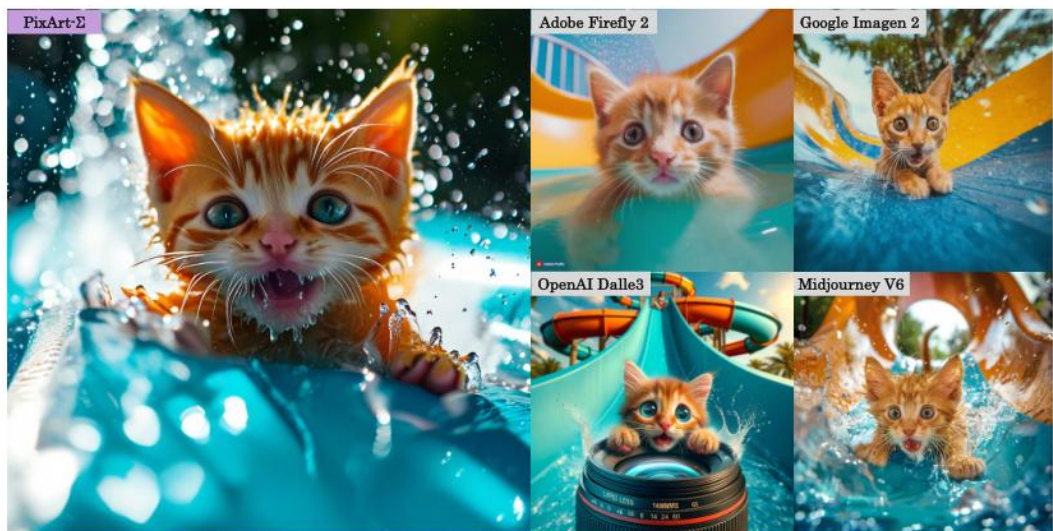
Prompt: 3D animation of a small, round, fluffy creature with big, expressive eyes explores a vibrant, enchanted forest. The creature, a whimsical blend of a rabbit and a squirrel, has soft blue fur and a bushy, striped tail. It hops along a sparkling stream, its eyes wide with wonder. Flowers that glow and change colors, trees with leaves in shades of purple and silver, and small floating lights that resemble fireflies. The creature stops to interact playfully with a group of tiny, fairy-like beings dancing around a mushroom ring.



Prompt: An extreme close-up of an gray-haired man with a beard in his 60s, he is deep in thought pondering the history of the universe as he sits at a cafe in Paris, his eyes focus on people offscreen as they walk as he sits mostly motionless, he is dressed in a wool coat suit coat with a button-down shirt, he wears a brown beret and glasses and has a very professorial appearance.

图 10: 将 PixArt- $\Sigma$  与其他四款文本到图像（T2I）产品进行比较：Firefly 2、Imagen 2、Dalle 3 和 Midjourney 6。由 PixArt- $\Sigma$  生成的图像与这些商业产品具有很强的竞争力。





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, toyon, captivating

图 11: 将 PixArt-Σ 与其他四款文本到图像 (T2I) 产品进行比较: Firefly 2、Imagen 2、Dalle 3 和 Midjourney 6。由 PixArt-Σ 生成的图像与这些商业产品具有很强的竞争力。



Isometric style farmhouse from RPG game, unreal engine, vibrant, beautiful, crisp, detailed, ultra detailed, intricate.



A car made out of vegetables how world looks like in 100 years, intricate, detailed



Chinese architecture, ancient style, mountain, bird, lotus, pond, big tree, 4K Unity, octane rendering.



A realistic landscape shot of the Northern Lights dancing over a snowy mountain range in Iceland, with long exposure to capture the motion and vibrant colors.



Game-Art - An island with different geographical properties and multiple small cities floating in space



Da Vinci's Last Supper oil painting in the style of Van Gogh



图 12: 由 PixArt-Σ 生成的高质量图像示例。PixArt-Σ 能够生成具有细致精细细节的高质量图像, 并且能够生成具有不同纵横比的多样化图像。



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



**图 13:** 由 PixArt- $\Sigma$  生成的高分辨率（4K）图像。PixArt- $\Sigma$  能够直接生成高质量的 4K HD（3840×2560）图像，同时保留细致的精细细节。

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



**图 14:** 由 PixArt- $\Sigma$  生成的高分辨率 (4K) 图像。PixArt- $\Sigma$  能够直接生成高质量的 4K HD (3840×2560) 图像, 同时保留细致的精细细节。