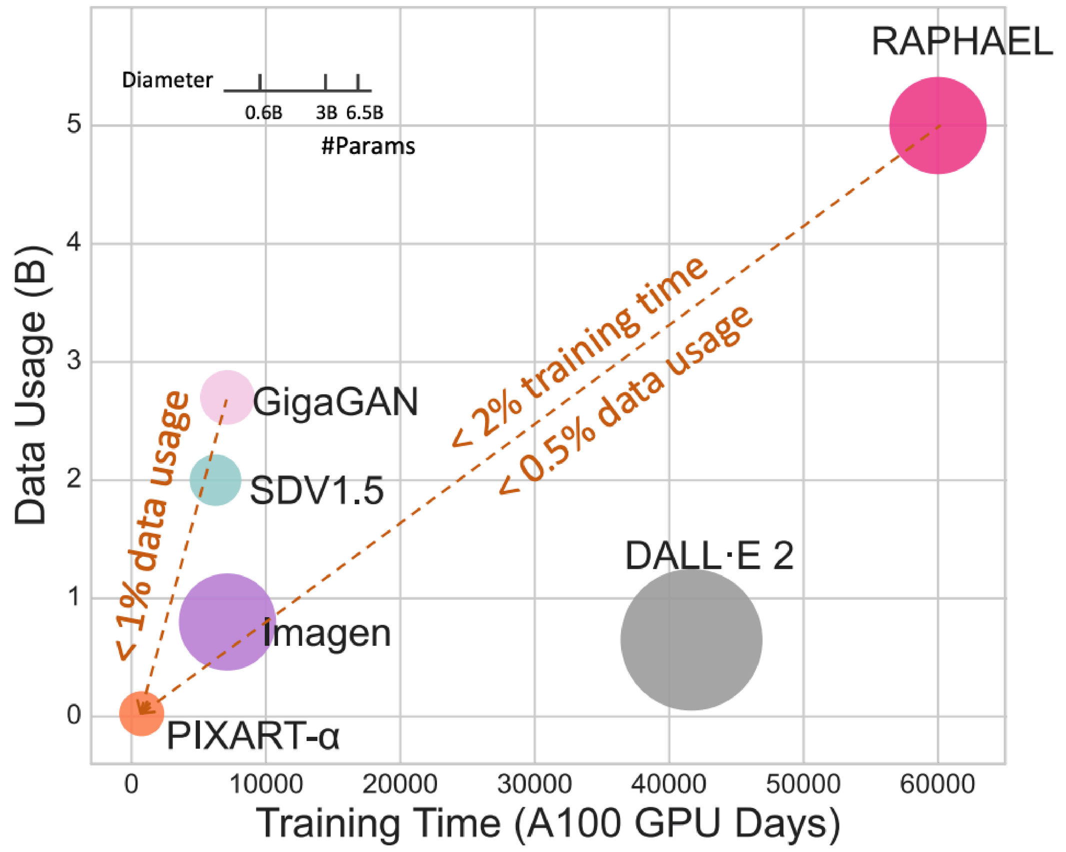
# PIXART-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis

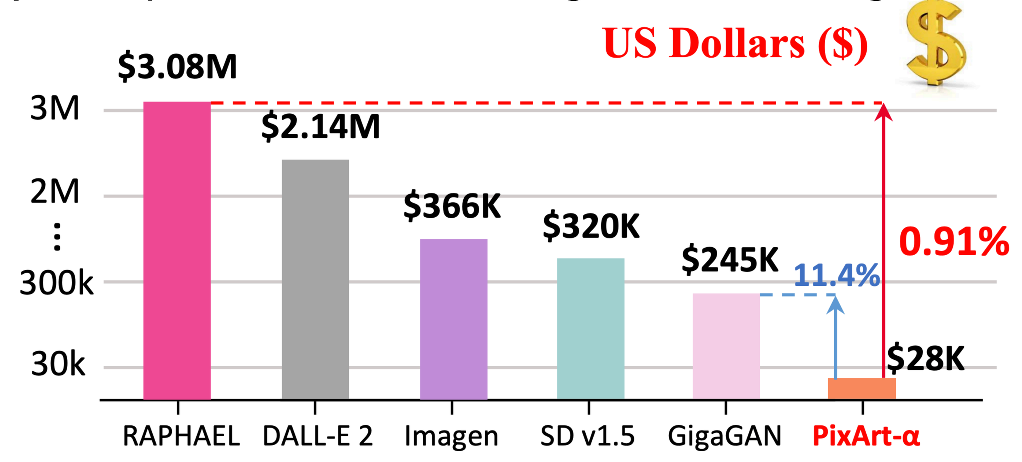
Junsong Chen[1,2,3*], Jincheng Yu[1,4*], Chongjian Ge[1,3*], Leiwei yao[1,4*], Enze Xie[2 †], Yue wu[1], Zhongdao Wang[1], James Kwok[4], Ping Luo[3], Huchuan Lu[2], Zhenguo Li[1]

[1]Huawei Noah's Ark Lab [2]DLUT [3]HKU [4]HKUST

## Problem Statement




(a) Comparison of data usage and training time



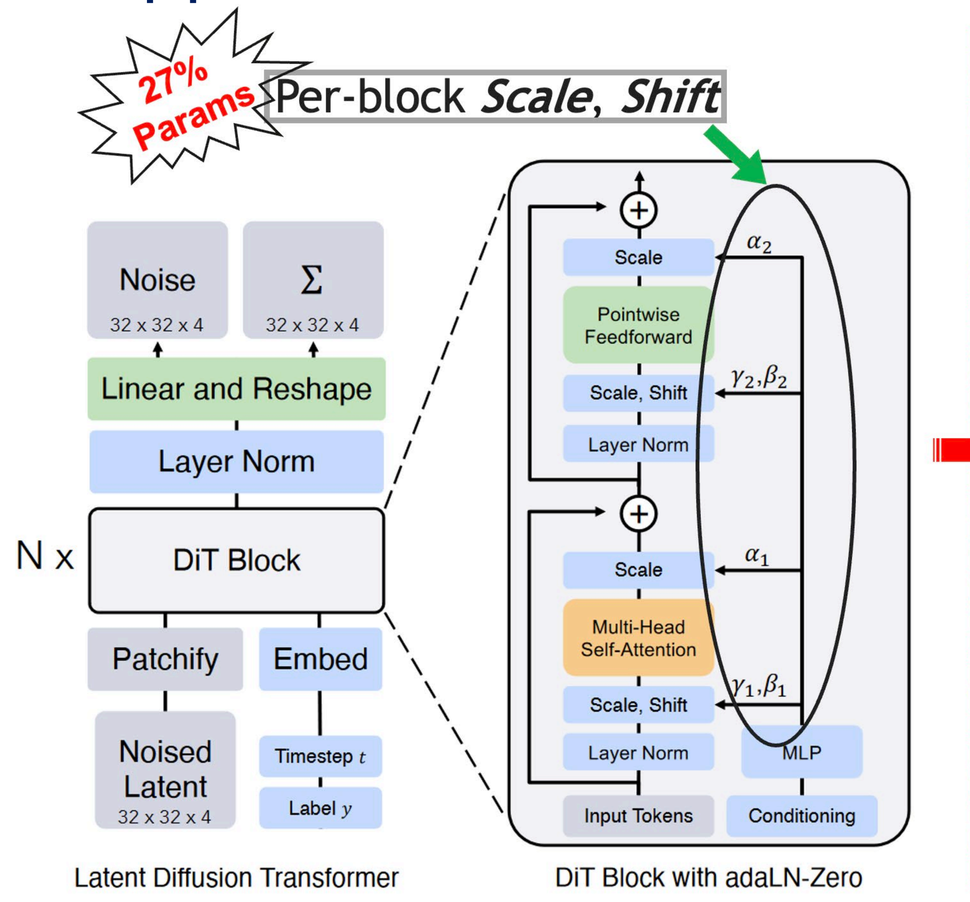

(b) Comparison of $CO_2$ emission and training cost

The AI Generative Content (AIGC) community faces a significant challenge as the most advanced Text-to-Image (T2I) models demand **enormous training costs**, equivalent to **millions of GPU hours**.

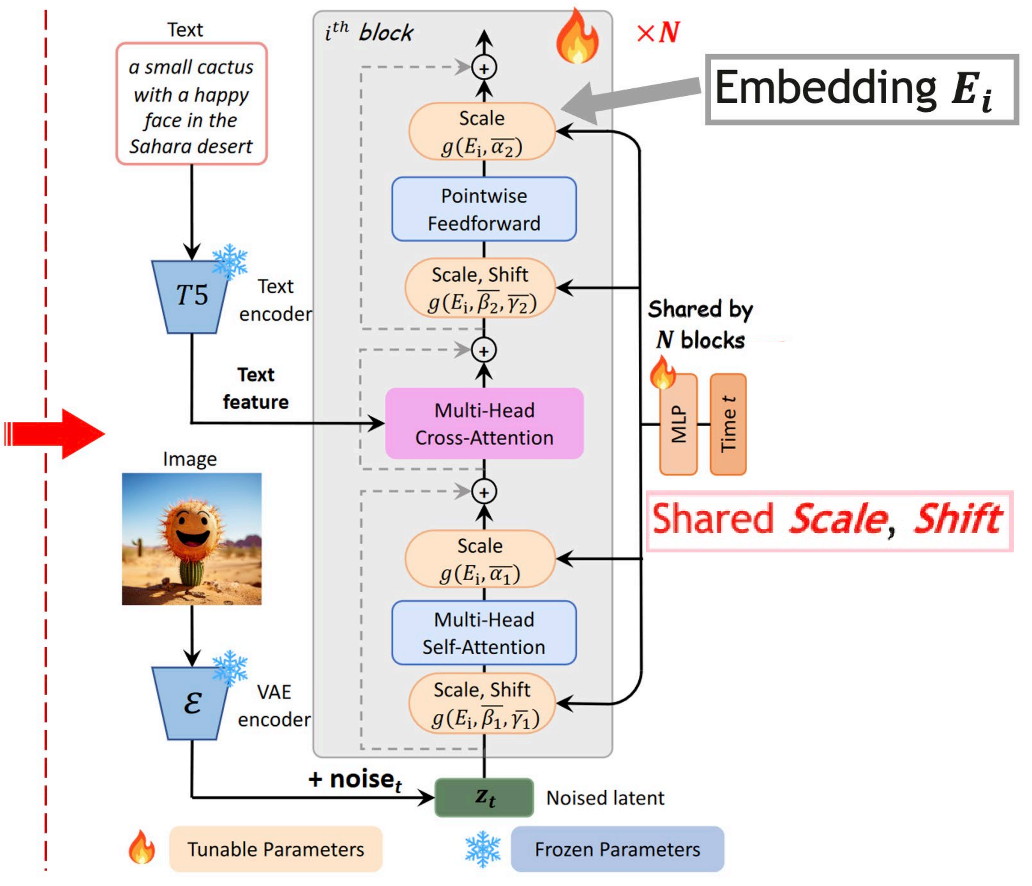
## Contributions

- We **decompose** the intricate text-to-image generation task into three streamlined subtasks.
- We introduce an **efficient Diffusion-Transformer** structure to fast adapt from class-conditioned DiT to text-conditioned PixArt-α.
- We propose an **auto-labeling pipeline** utilizing the state-of-the-art vision-language model to generate captions on the SAM.

## Our Approach

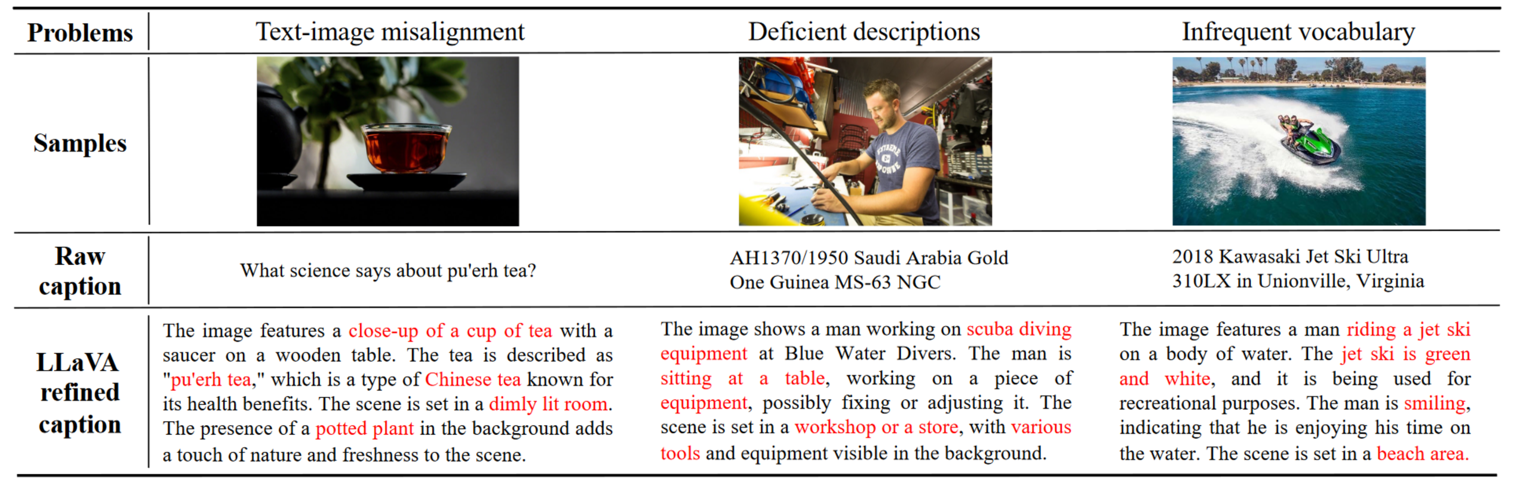



Original DIT Architecture




Efficient T2I Transformer

Model architecture of PIXART-α. A cross-attention module is integrated into each block to inject textual conditions. To optimize efficiency, all blocks share the same adaLN-single parameters for time conditions.




Low density

| Dataset | VN/DN | Total Noun | Average |
|---|---|---|---|
| LAION | 210K/2461K = 8.5% | 72.0M | 6.4/Img |
| LAION-LLaVA | 85K/646K = 13.3% | 233.9M | 20.9/Img |
| SAM-LLaVA | 23K/124K = 18.6% | 327.9M | 29.3/Img |
| Internal | 152K/582K = 26.1% | 136.6M | 12.2/Img |

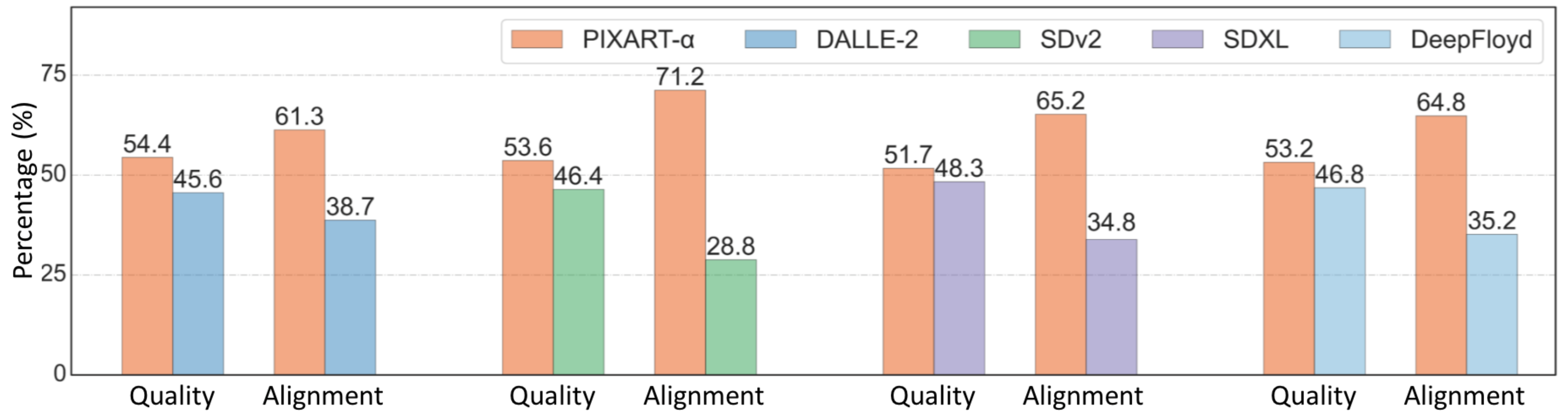
High density

Statistics of noun concepts for different datasets.

LAION raw captions v.s LLaVA refined captions. LLaVA provides high-information density captions that aid the model in grasping more concepts per iteration and boost text-image alignment efficiency.

## Appealing Generations


## Quantitative Experiments




User study on 300 fixed prompts from Ernie-vilg 2.0

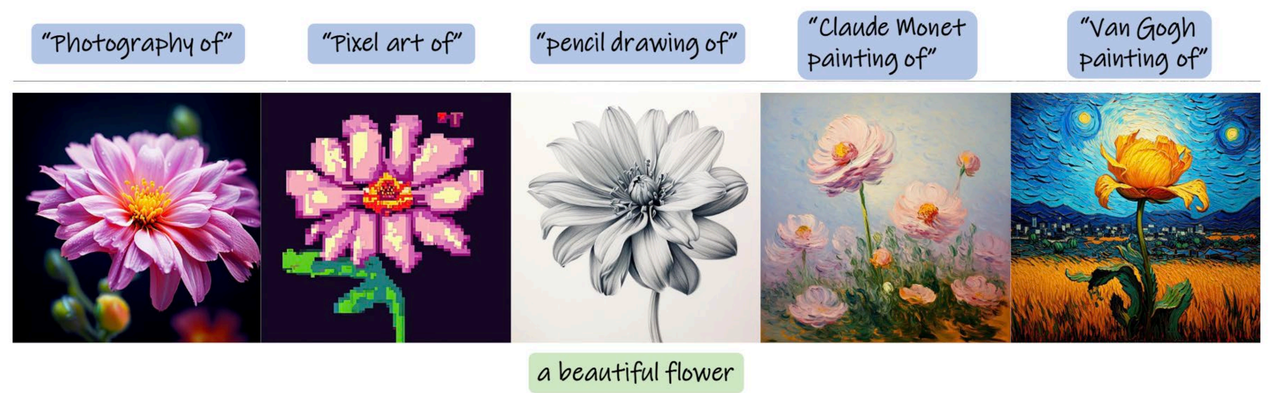
| Method | Type | #Params | #Images | FID-30K↓ | GPU days |
|---|---|---|---|---|---|
| DALL·E | Diff | 12.0B | 250M | 27.50 | - |
| GLIDE | Diff | 5.0B | 250M | 12.24 | - |
| LDM | Diff | 1.4B | 400M | 12.64 | - |
| DALL·E 2 | Diff | 6.5B | 650M | 10.39 | 41,667 A100 |
| SDv1.5 | Diff | 0.9B | 2000M | 9.62 | 6,250 A100 |
| GigaGAN | GAN | 0.9B | 2700M | 9.09 | 4,783 A100 |
| Imagen | Diff | 3.0B | 860M | 7.27 | 7,132 A100 |
| RAPHAEL | Diff | 3.0B | 5000M+ | 6.61 | 60,000 A100 |
| PIXART-α | Diff | 0.6B | 25M | 7.32 | 753 A100 |

We thoroughly compare the PIXART-α with recent T2I models

| Model | Attribute Binding | | | Object Relationship | | Complex↑ |
|---|---|---|---|---|---|---|
| | Color ↑ | Shape↑ | Texture↑ | Spatial↑ | Non-Spatial↑ | |
| Stable v1.4 | 0.3765 | 0.3576 | 0.4156 | 0.1246 | 0.3079 | 0.3080 |
| Stable v2 | 0.5065 | 0.4221 | 0.4922 | 0.1342 | 0.3096 | 0.3386 |
| Composable v2 | 0.4063 | 0.3299 | 0.3645 | 0.0800 | 0.2980 | 0.2898 |
| Structured v2 | 0.4990 | 0.4218 | 0.4900 | 0.1386 | 0.3111 | 0.3355 |
| Attn-Exct v2 | 0.6400 | 0.4517 | 0.5963 | 0.1455 | 0.3109 | 0.3401 |
| GORS | 0.6603 | 0.4785 | 0.6287 | 0.1815 | 0.3193 | 0.3328 |
| Dalle-2 | 0.5750 | 0.5464 | 0.6374 | 0.1283 | 0.3043 | 0.3696 |
| SDXL | 0.6369 | 0.5408 | 0.5637 | 0.2032 | 0.3110 | 0.4091 |
| PIXART-α | 0.6886 | 0.5582 | 0.7044 | 0.2082 | 0.3179 | 0.4117 |

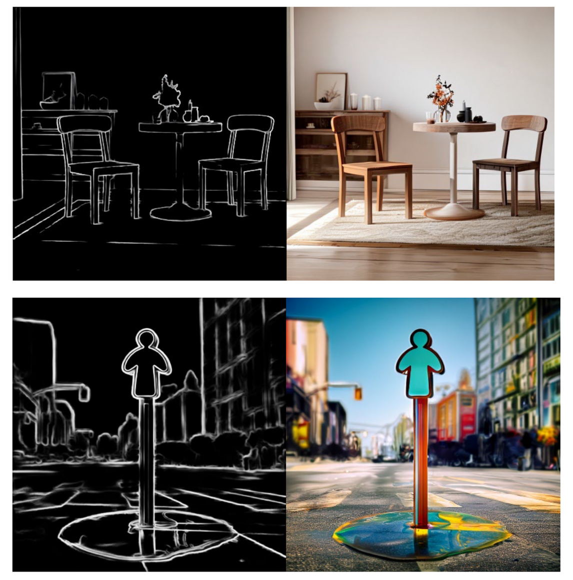
Alignment evaluation on T2I-CompBench.

## Generalization Extensions

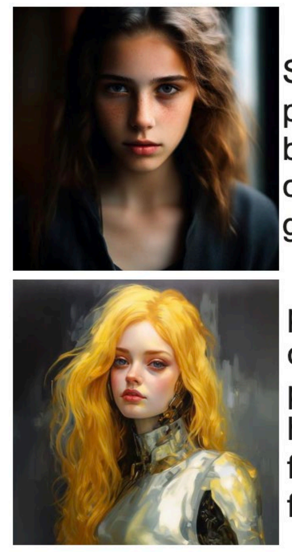



Style control with text.


PixArt-Dreambooth



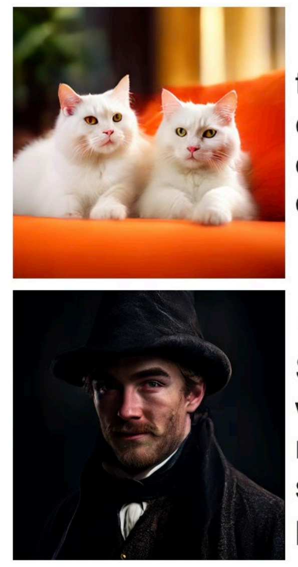
PixArt-ControlNet



Self-portrait oil painting, a beautiful cyborg with golden hair, 8k

portrait photo of a girl, photograph, highly detailed face, depth of field

PixArt-LCM (4 steps)
1024px 0.51s



two white cats playing on top of the orange sofa

19th century Scottish wizard with a mysterious smile and a piercing gaze

PixArt-DMD (1 step)
512px 0.1s