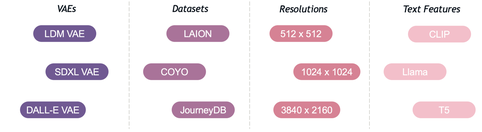
PIXART-Σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation

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

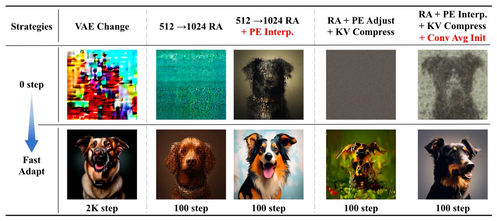
[1]Huawei Noah's Ark Lab [2]DLUT [3]HKU [4]HKUST

Problem Statement

Parts will be Continuously updating:



(a) Multiple elements we may change in development



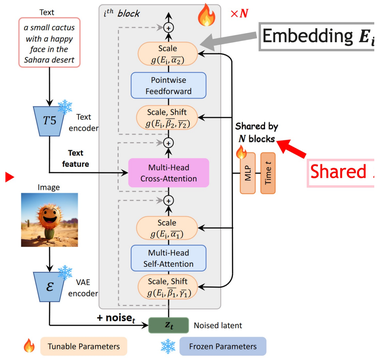
(b) **Weak-to-Strong Training Strategy**

PIXART-Σ is capable of directly generating images at 4K resolution. it evolves from the 'weaker' baseline to a 'stronger' model via incorporating higher quality data, a process we term "weak-to-strong training".

Contributions

- We collect a high-quality dataset superior to that used in PixArt-α, focusing on **High-quality images** and **Dense and accurate captions**.
- We introduce an **Efficient Token Compression** to effectively reduce training and inference time by ~34% for high-resolution image generation.
- We propose **Weak-to-Strong Training Strategy** utilizing several fine-tuning techniques to rapidly adapt from a weak model to a strong model efficiently

Our Approach



Efficient T2I Transformer

Model architecture of PIXART-Σ. We merge KV tokens in spatial space in self-attention block to reduce the computation complexity, specially for high-resolution image generation.

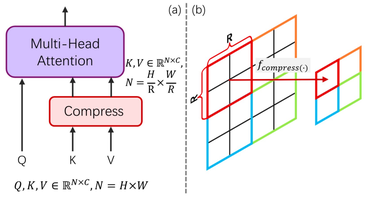
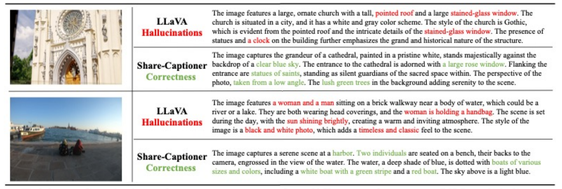


Fig. 7: Design of KV Token Compression. We merge KV tokens in spatial space



Comparative illustration of hallucinations: Contrasting differences in hallucination occurrences between LLaVA and Share-Captioner, with red indicating hallucinations and green denoting correctness.

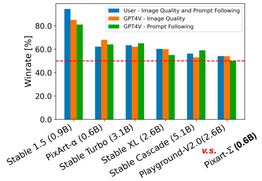
Table 1: Statistics of noun concepts for different datasets. VN: valid distinct nouns (appearing more than 10 times); **DN**: total distinct nouns; **Average**: average noun count per image; **ACL**: Average Caption length.

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

Appealing Generations

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.*

Quantitative Experiments



We compare the SOTAs with User preference and using GPT4v for image quality and instruct following ability.

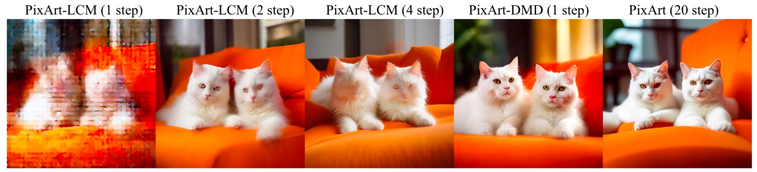
GPV4 prompts for evaluation

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.

We create a 30K PixEval dataset for High quality FID and CLIP-Score assessment

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-α	0.6	8.65	0.2787
PixArt-Σ	0.6	8.23	0.2797

Generalization Extensions



Prompt: two white cats playing on top of the orange sofa, very comfortable Best quality.

PixArt-DMD 1 step generation

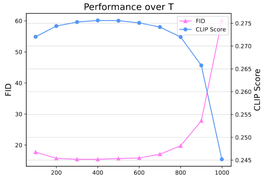


Fig. 10: Base model + DMD performance over T.

Table 4: 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

Timestep T need to be set to T=400 during training than T = 999 for better convergence

Training Details

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