

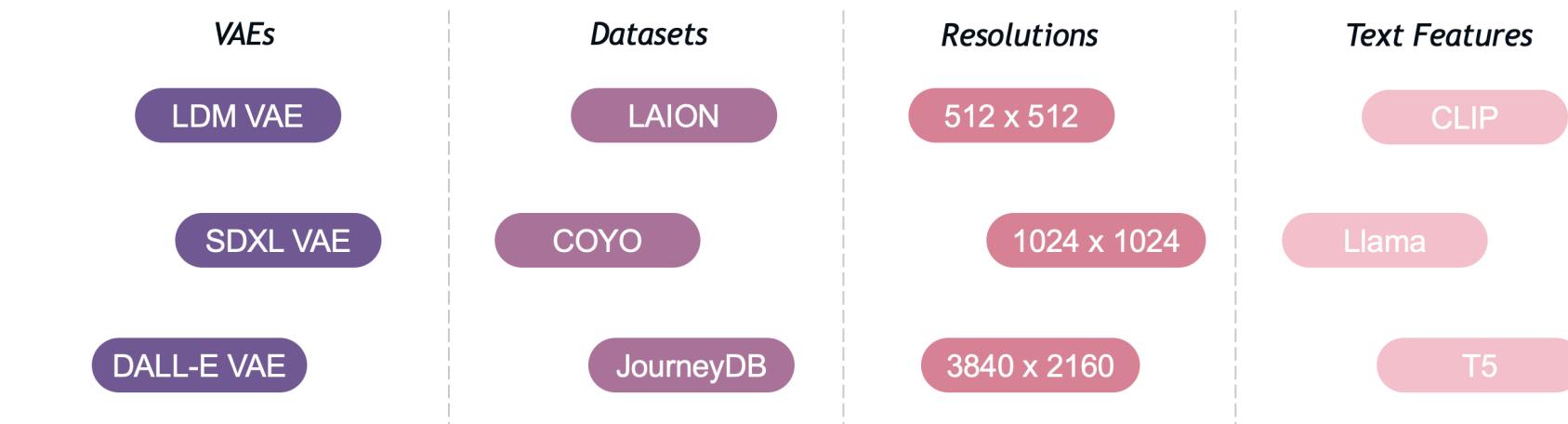
PIXART-Σ: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation

Junsong Chen^{1,2,3*}, Chongjian Ge^{1,3*}, Enze Xie^{2†}, Yue wu¹, Leiwei yao^{1,4*}, Xiaozhe Ren¹, Zhongdao Wang¹, James Kwok⁴, Ping Luo³, Huchuan Lu², Zhenguo Li¹ ¹Huawei Noah's Ark Lab ²DLUT ³HKU ⁴HKUST

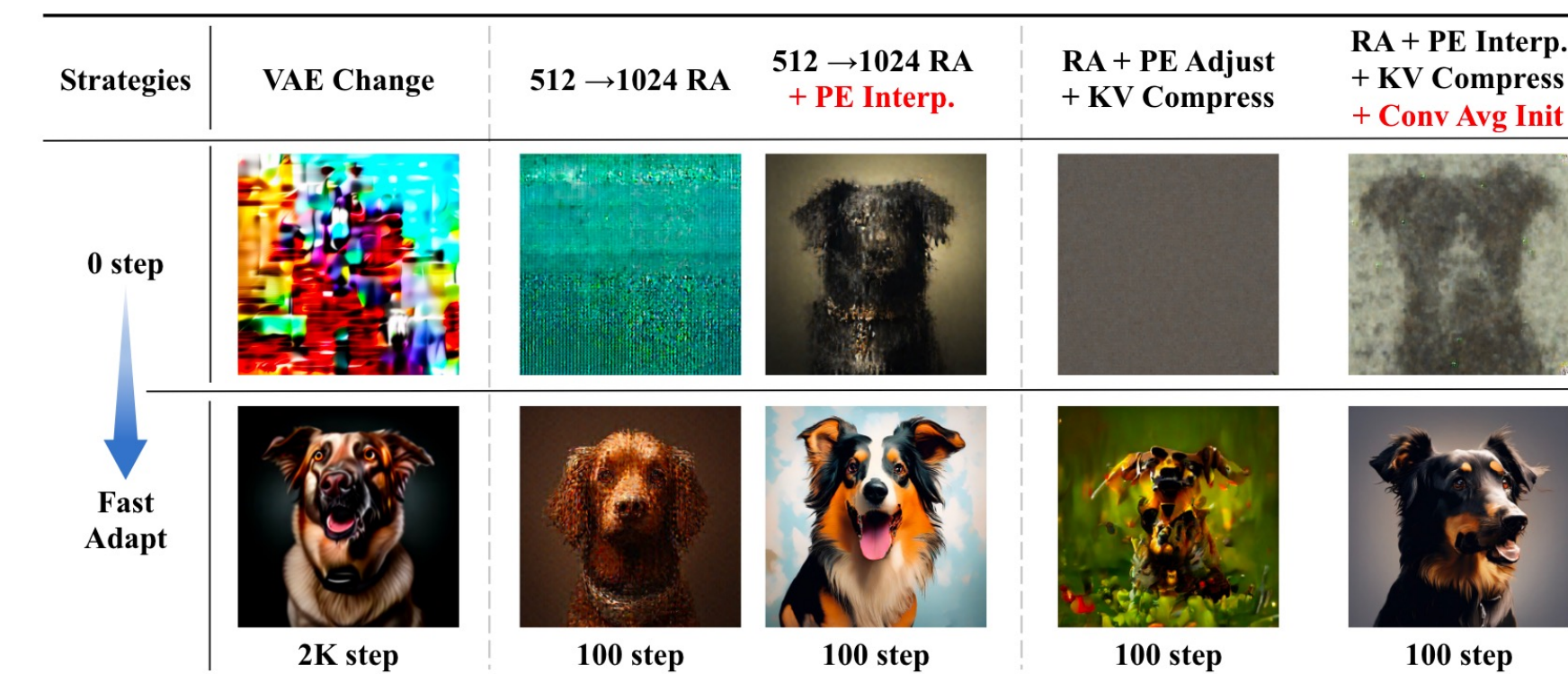


Problem Statement

Parts will be Continuously updating:



(a) Multiple elements we may change in development



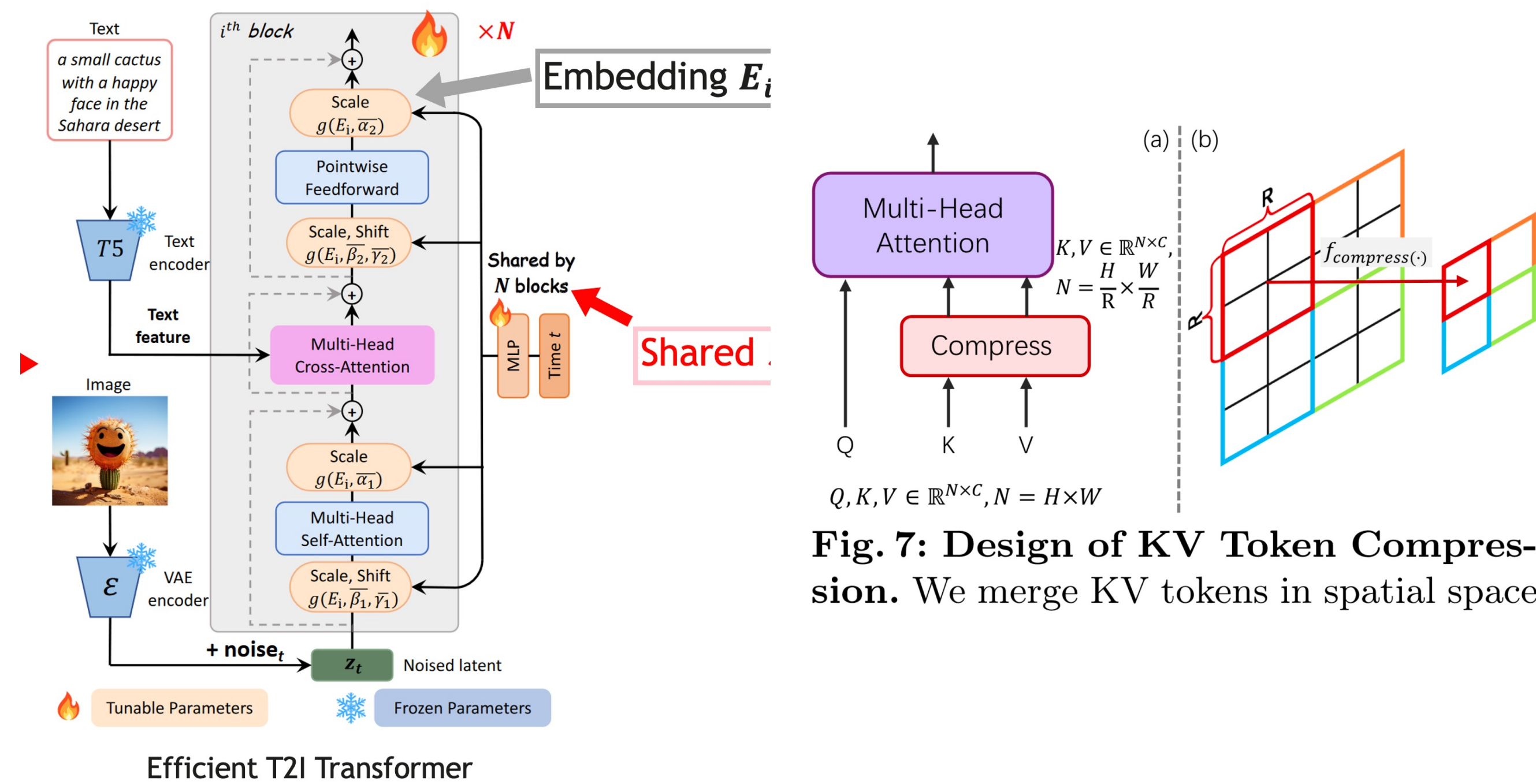
(b) Weak-to-Strong Training Strategy

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



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

| Image | LLaVA Hallucinations | Share-Captioner Correctness |
|-------|---|--|
| | The image features a large, ornate church with a tall, pointed roof and a large stained-glass window. 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 stained-glass window. The presence of statues and a clock on the building further emphasizes the grand and historical nature of the structure. | The image captures the grandeur of a cathedral, painted in a pristine white, stands majestically against the backdrop of a clear blue sky. The entrance to the cathedral is adorned with a large rose window. Flanking the entrance are statues of saints, standing as silent guardians of the sacred space within. The perspective of the photo, taken from a low angle, The lush green trees in the background adding serenity to the scene. |
| | The image features a woman and a man 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 woman is holding a handbag. The scene is set during the day, with the sun shining brightly, creating a warm and inviting atmosphere. The style of the image is a black and white photo, which adds a timeless and classic feel to the scene. | The image captures a serene scene at a harbor. Two individuals 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 boats of various sizes and colors, including a white boat with a green stripe and a red boat. The sky above is a light blue. |

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.

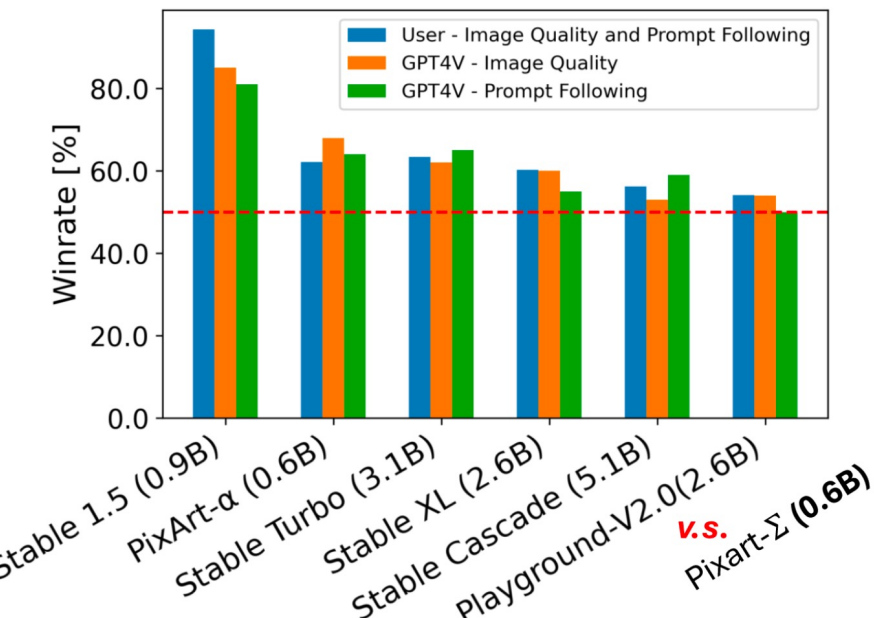
| Dataset | Volume | Caption | VN/DN | Total Noun | ACL | Average |
|------------|--------|-----------------|------------|------------|-----|----------|
| Internal-α | 14M | Raw | 187K/931K | 175M | 25 | 11.7/Img |
| Internal-α | 14M | LLaVA | 28K/215K | 536M | 98 | 29.3/Img |
| Internal-α | 14M | Share-Captioner | 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

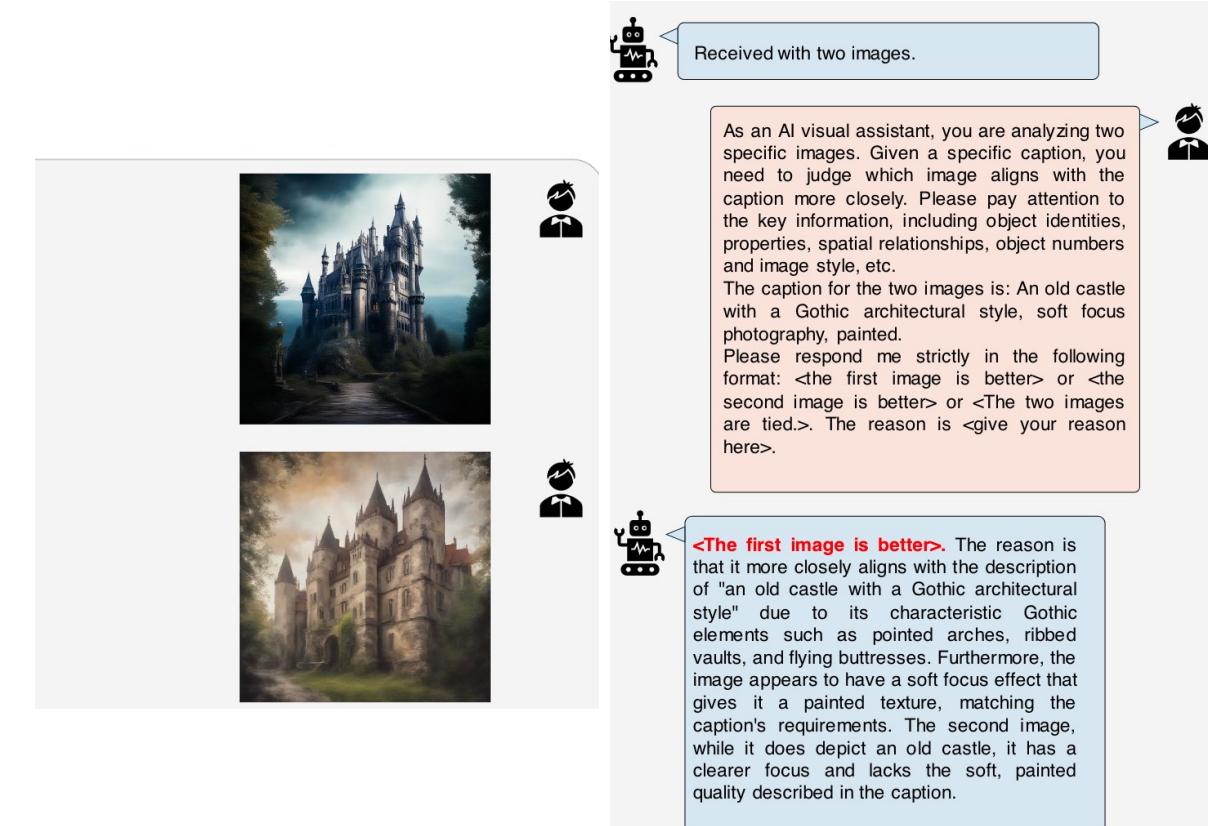
A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, 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.



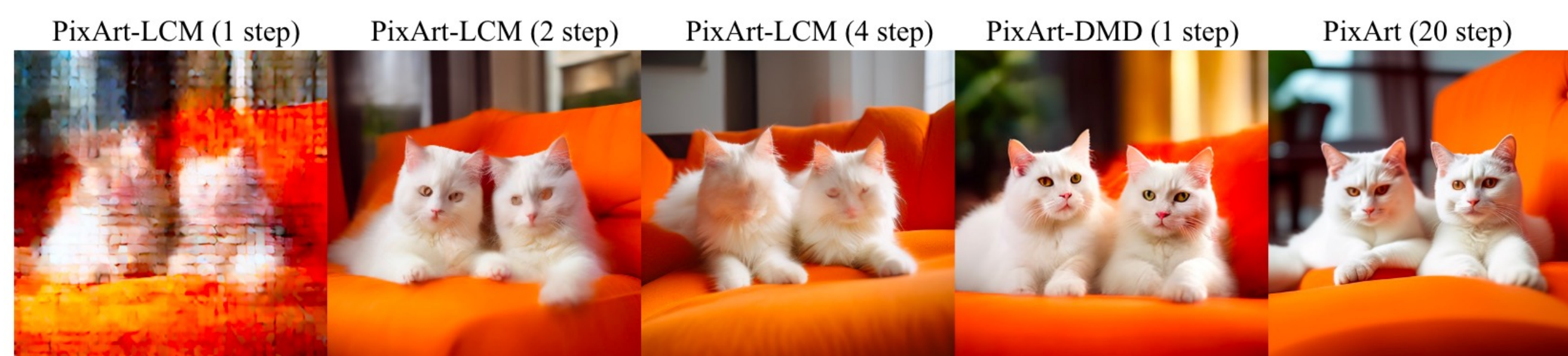
GPV4 prompts for evaluation

| 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

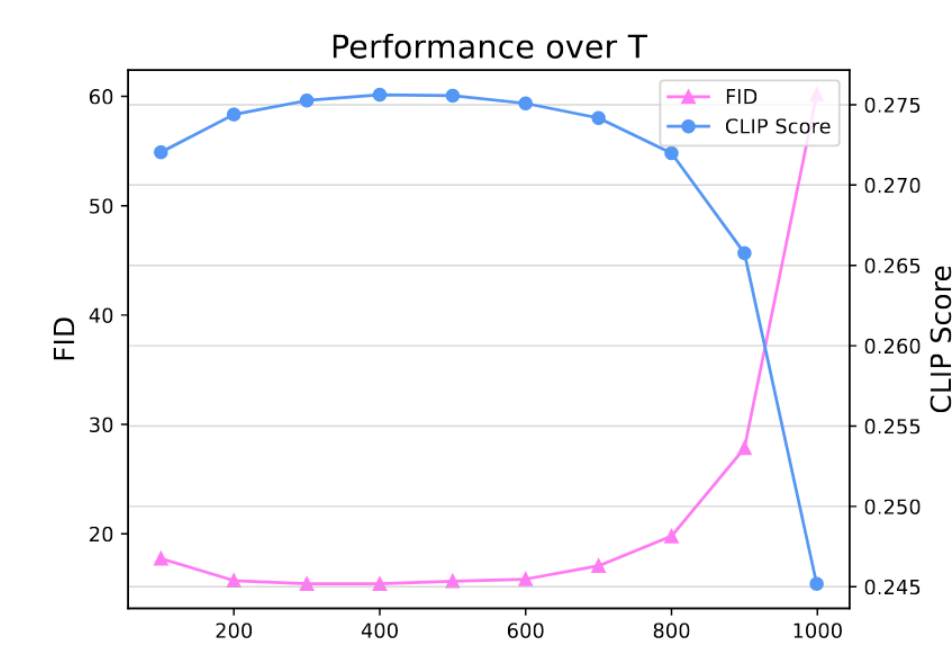


Fig. 10: Base model + DMD performance over T.

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

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 |

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 |