ArXiv '24

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

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, Zhenguo Li

Jeonghoon Park, happypjh2001@unist.ac.kr
Undergraduate Research Intern,
Ubiquitous Artificial Intelligence Lab,
Department of Computer Science and Engineering,
Ulsan National Institute of Science and Technology

Reason to choose this paper

For research…

MobileDiffusion

- ◆ MobileDiffusion: Instant Text-to-Image Generation on Mobile Devices
- ◆ 512 x 512 Resolution

On-device + 4K diffusion

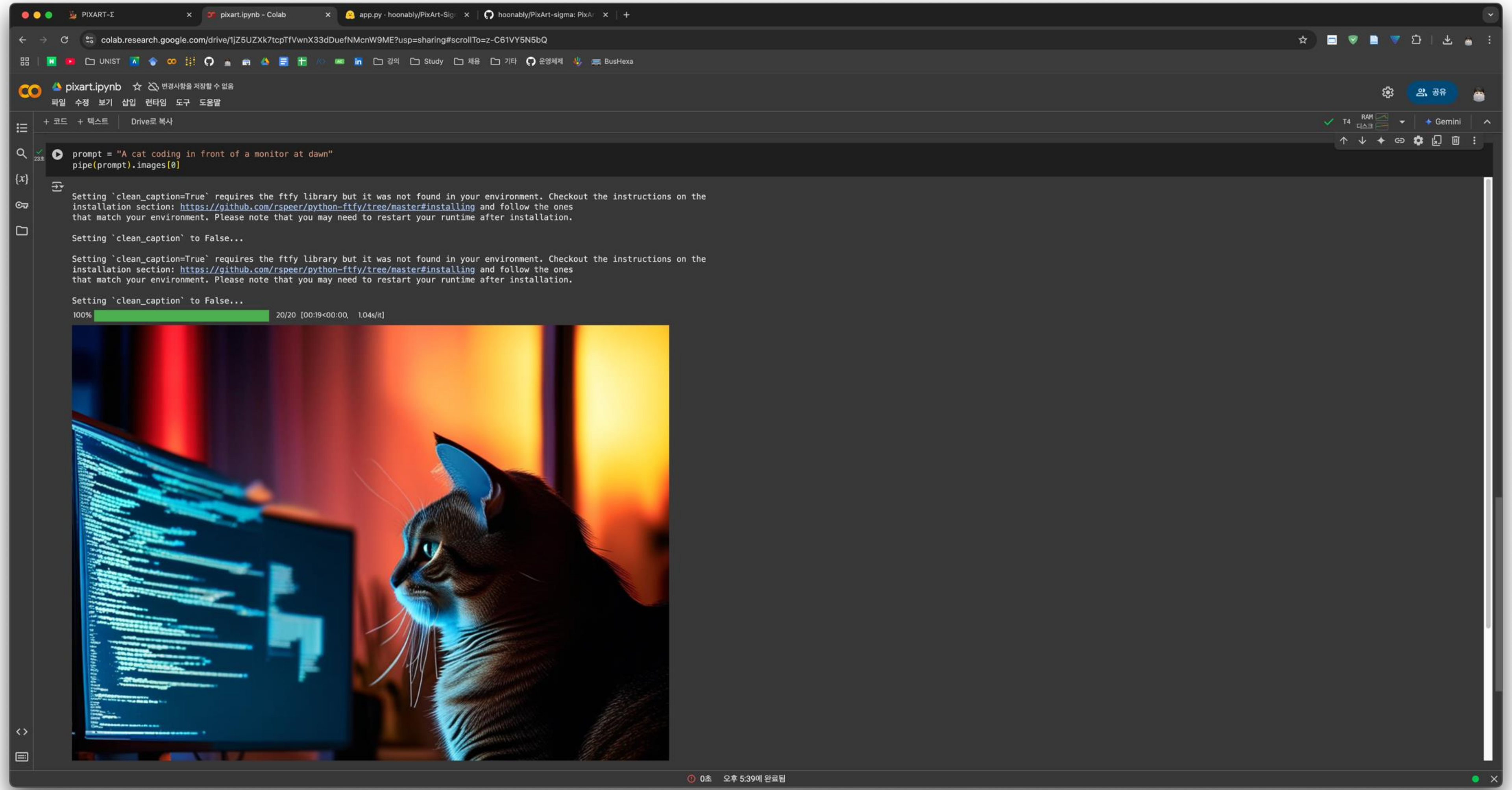
- ◆ No papers on implementing 4K at on-device
- ◆ Decide on a topic for my research

4K diffusion paper

- ◆ Exploring 4K diffusion technology
- ◆ Latest paper

Preview : Result of a run in Colab T4

PixArtAlphaPipeline.from_pretrained("PixArt-alpha/PixArt-Sigma-XL-2-1024-MS", torch_dtype=torch.float16)



Previous works

PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (ICLR, 2024 Spotlight)

Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.

➔ First Transformer-based Diffusion Model (DiT) capable of generating up to 1024×1024 resolution

Stable Diffusion : High-resolution image synthesis with latent diffusion models (CVPR, 2022)

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.

➔ Utilizes the Latent Diffusion Model (LDM) structure to generate high-resolution images beyond 1024×1024

DALL-E 3 (OpenAI, 2023)

➔ Utilizes GPT-4-based text comprehension to more accurately reflect prompts

Data Analysis : Higher Aesthetic and higher Resolution

Effective training with limited data

	Data	Data Resolution
Internal-α	14M	Only 256 ~ 1K
Internal-Σ	**33M**	1K~4K (33M) real photo 4K (8M)
SD v1.5 (open-source)	2B	512x512, 768x768

Data Analysis : Higher Aesthetic and higher Resolution

However, well scored!

Models	#Params (B)	FID $\downarrow$	CLIP-Score $\uparrow$
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

Data Analysis : Enhanced caption accuracy

PixArt-α (LLaVa) -> certain hallucination problem

PixArt-Σ (Share-Captioner) -> generate detailed and correct captions -> augmenting the collected raw prompts

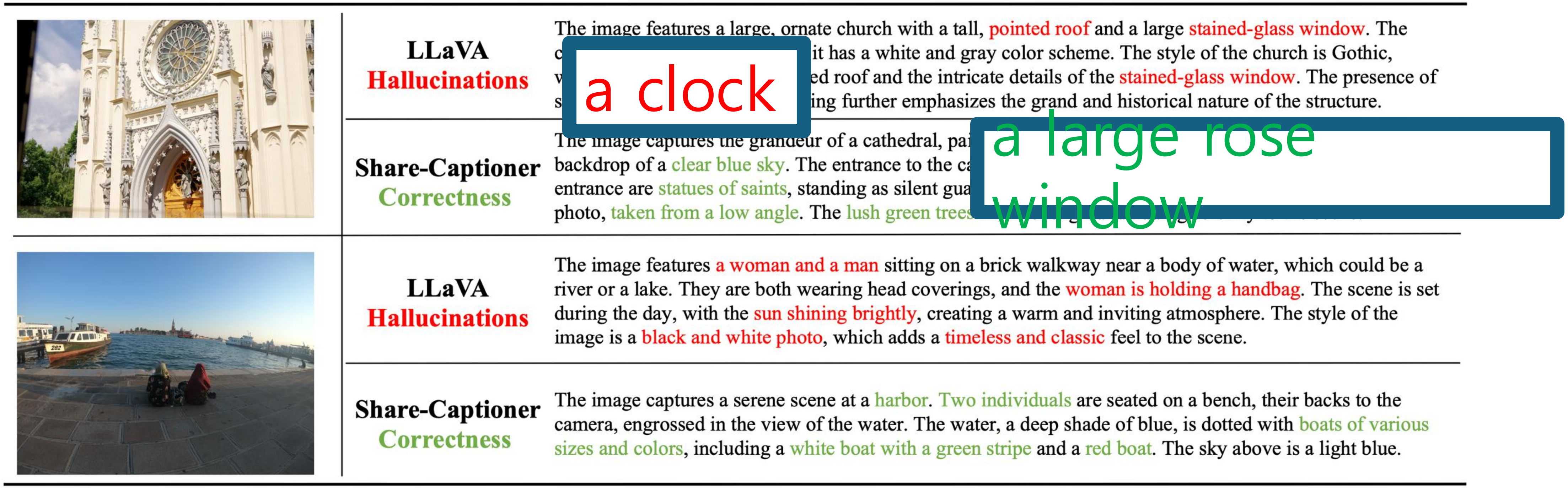


Fig. 5: Comparative illustration of hallucinations: Contrasting differences in hallucination occurrences between LLaVA and Share-Captioner, with red indicating hallucinations and green denoting correctness.

hallucination : 환각,착시

(Skip)

LLaVa

“Visual Instruction Tuning” (2023)

➔ a study to create a visual version (Vision-Language Model, VLM) of GPT-4.

- Based on: CLIP + LLaMA (Language Model)
- Purpose: Multimodal model to view images and perform “description, question-answer (Q&A), summarization, etc.”
- Features:
  - Utilizes CLIP to convert images into linguistic representations.
  - Large Language Model (LLaMA) to generate text.
  - Performs a similar role to the traditional GPT-4V.
  - However, it can be less accurate and potentially lacks fine-grained information.

(Skip)

Share-Captioner

“ShareGPT4V: Improving Large Multi-Modal Models with Better Captions” (2023)

➔ Share-Captioner is a model to overcome the limitations of LLaVA and generate more sophisticated captions.

- Based on: GPT-4V (GPT-4 with Vision)
- Purpose: Generate more accurate and detailed image captions
- Features:
  - Utilizes GPT-4V to generate more accurate and detailed descriptions.
  - Longer sentences, more detail than LLaVA.

Data Analysis : Increased caption length

Internal-α : <= 120 tokens

Internal-Σ : <= 300 tokens

Share-Captioner(60%) + raw(40%) -> reduce potential biases

(Not using raw data can introduce bias!)

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

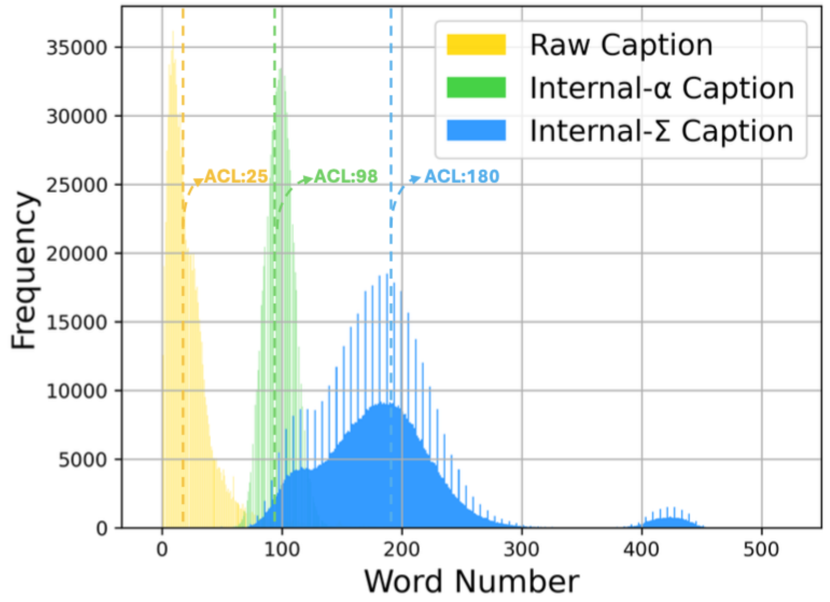


Fig. 6: Histogram Visualization of the Caption Length. We randomly select 1M captions from the raw captions, Internal-α, and Internal-Σ to draw the corresponding histogram. *ACL* denotes the average caption length.

Data Analysis : High-Quality Evaluation Dataset

SoTA : State of the Art

Most SoTA T2I models chose **MSCOCO** (MobileDiffusion too)

-> Not enough to evaluate aesthetics and text-image alignment

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. 12: 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.

This paper use 30,000 high quality dataset

Efficient DiT Design : Previous problems

in PixArt-α

Self-Attention computation increases proportional to **the square of the number of tokens**

➔ **O(N²)**

4K resolution needs higher number of tokens → **Model execution = slow**

Memory usage spikes when generating 4K → **GPU costs = increase**

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

Efficient DiT Design : Key-Value Token Compression

- Compress Key (K) and Value (V) using **Group Convolution on Stride 2**

➔ reduce the number of tokens by **N → N/R^2**

- Using **1<=R<=4** without losing too much accuracy

➔ reduce computation by about 34%

$$\text{from } O(N^2) \text{ to } O\left(\frac{N^2}{R^2}\right)$$

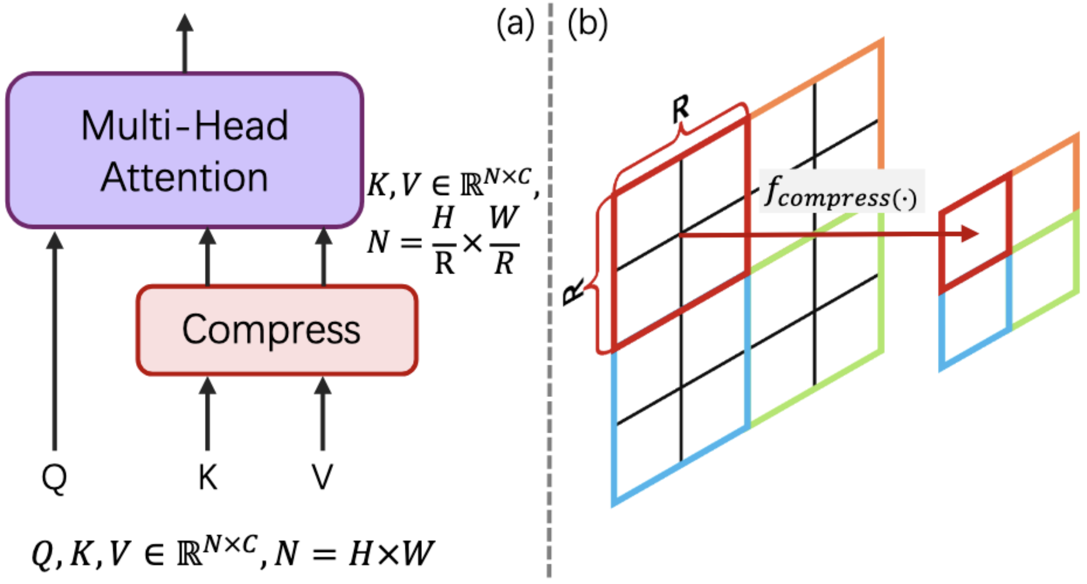


Fig. 7: Design of KV Token Compression. We merge KV tokens in spatial space to reduce the computation complexity.

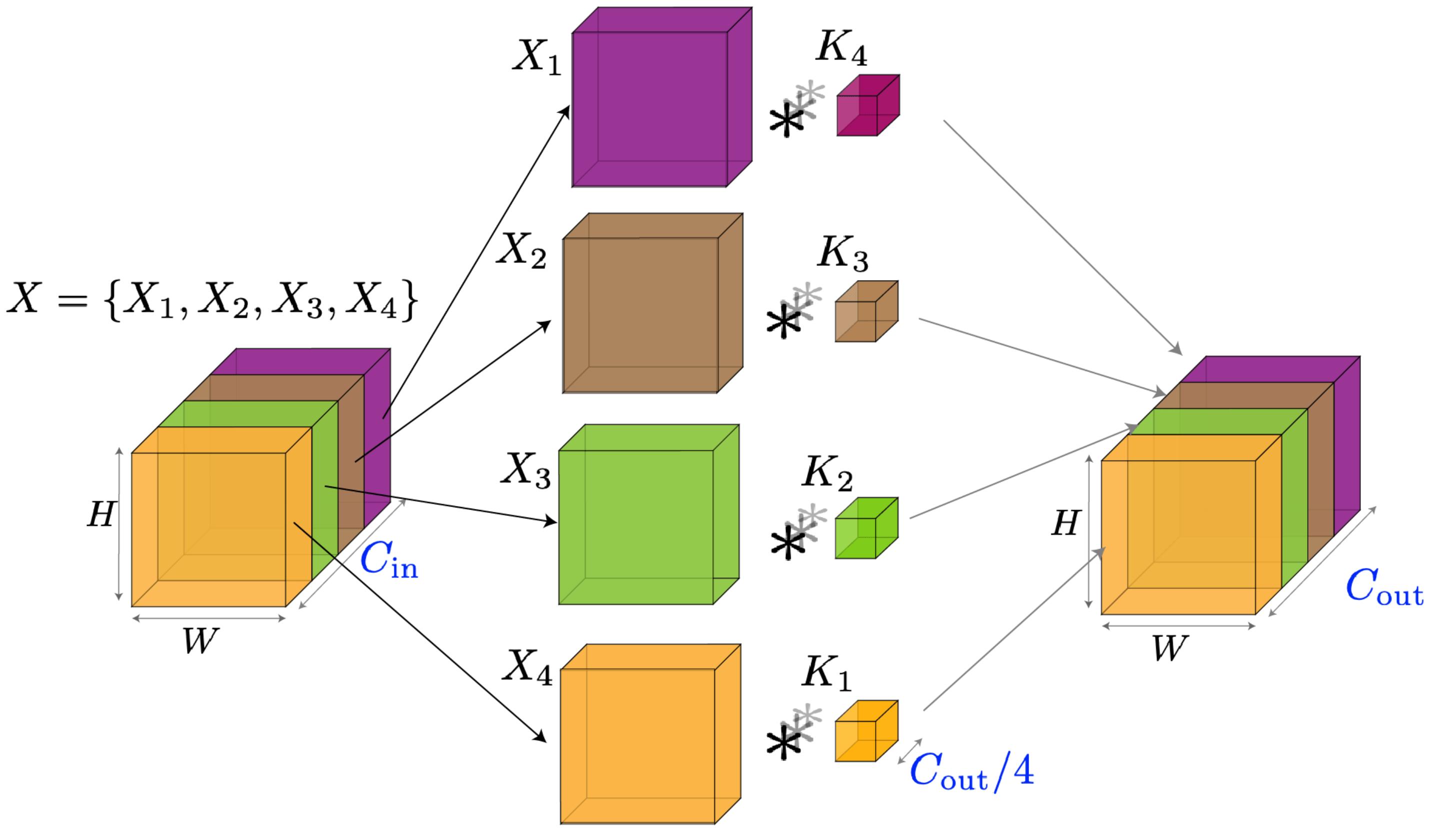
(Skip)

Group Convolution?

Divide input channels into groups and perform convolution on each independently

➔ Less computation

➔ More speed



(Skip)

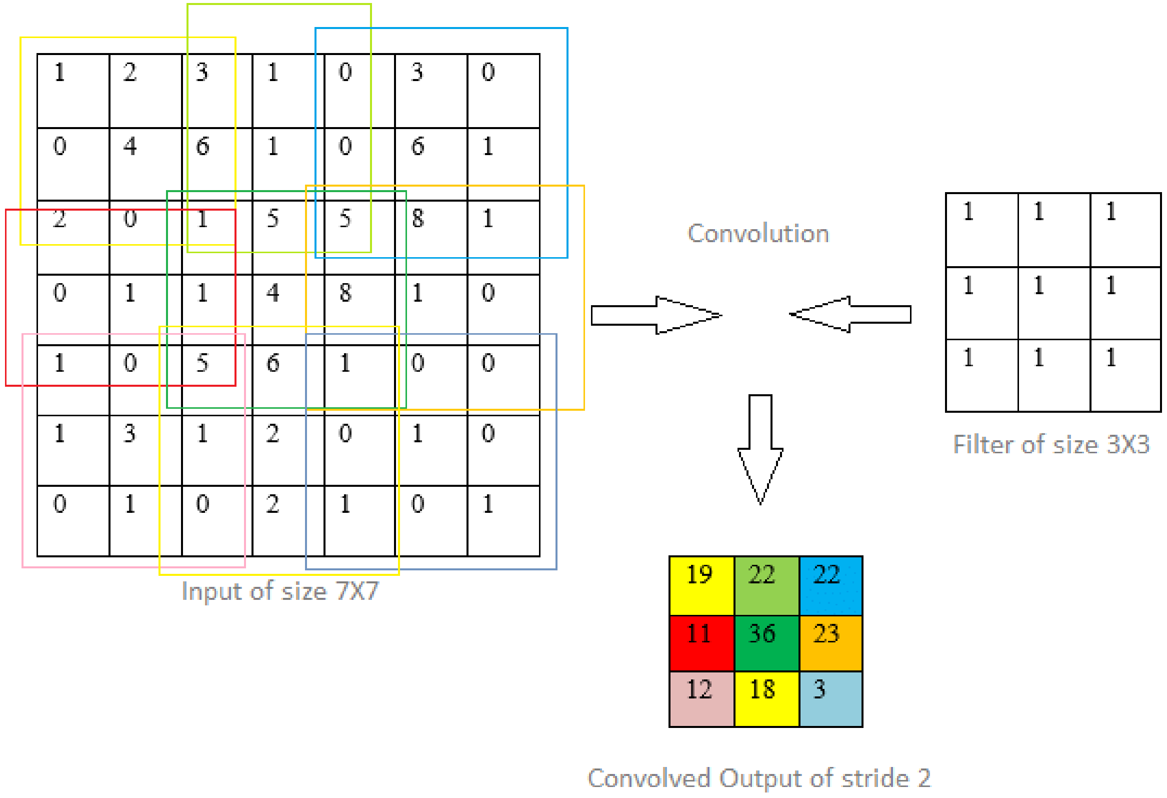
Stride 2?

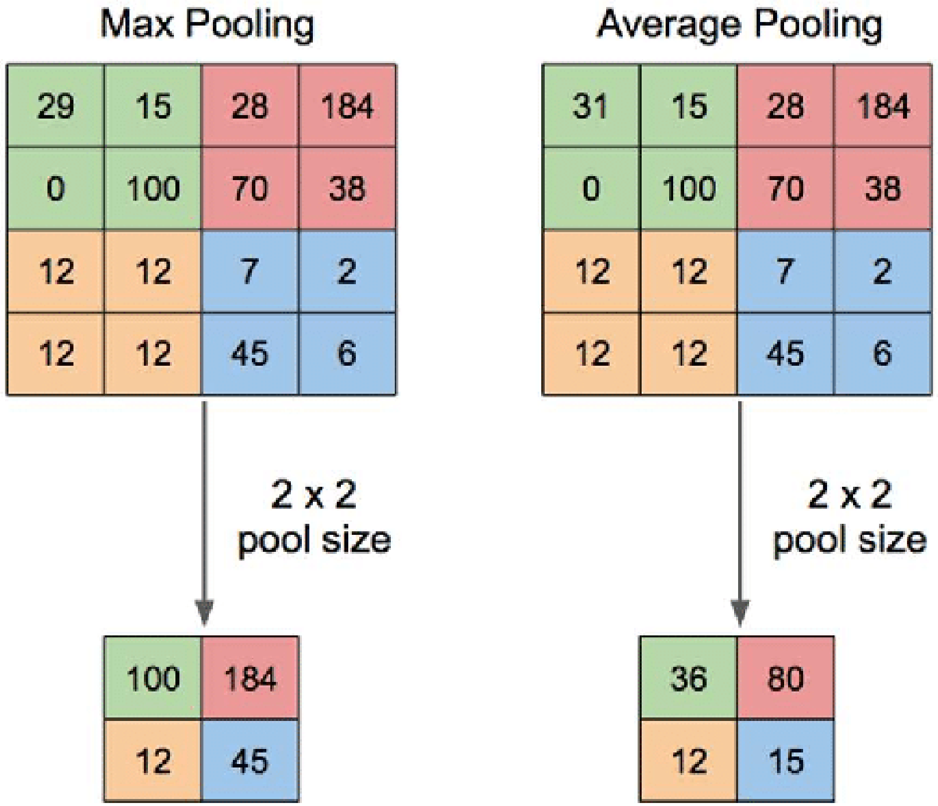
Moves 2 pixels at a time, so the output is half the size of the input

➔ Downsampling

➔ Less computation

➔ Keep important information than Max Pooling or Average Pooling(general downsampling)





Efficient DiT Design : Key-Value Token Compression

Effect of Compression Ratio

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 rations 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.

Table 3: KV-Token Compression Settings in Image Generation. This study employs FID, CMMD, and CLIP-Score metrics to assess the impact of various token compression components, such as compression ratio, positions, operators, and varying resolutions. Speed calculation in Tab. 3c is Second/Iteration/384 Batch-size.

Weak-to-Strong Training Strategy

Adapting model to new VAEs

◆ PixArt-α : VAE (8x downsampling)

➔ PixArt-Σ : **Stable Diffusion XL(SDXL) VAE (4x downsampling)**

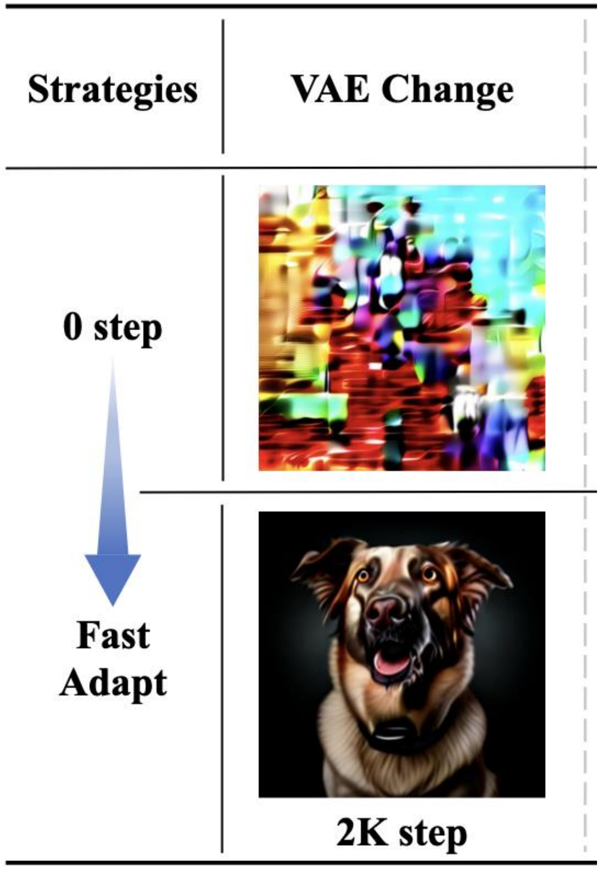
➔ Preserve details

◆ If training T2I models from scratch = resource-intensive

➔ choosing fine-tuning

◆ How?

➔ fine-tuning quickly converges at **2K training steps**



(a)

Weak-to-Strong Training Strategy

Adapting to Higher-Resolution

fine-tune from a low-resolution (LR) model to a high-resolution (HR) moc

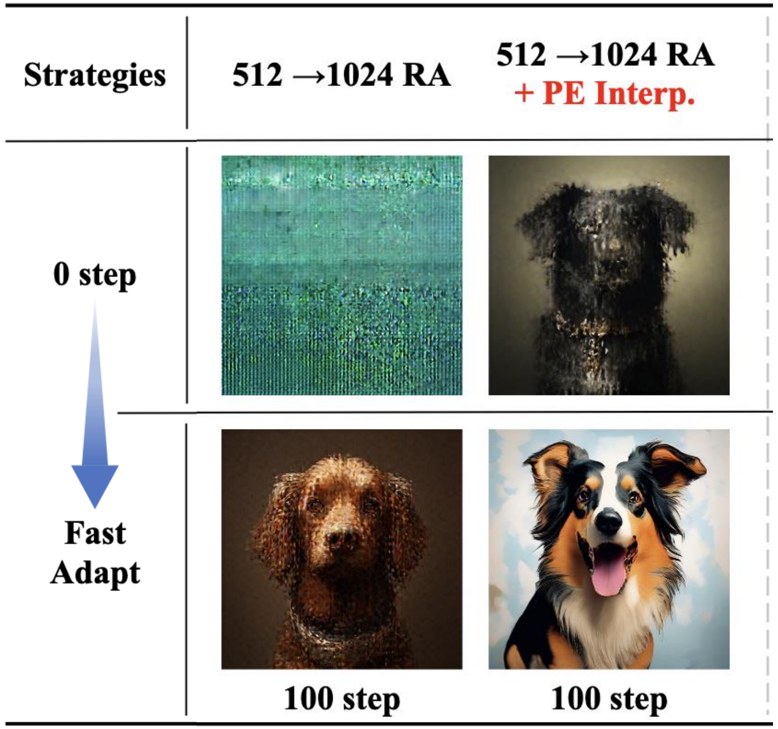
➔ observe a performance degradation 😢

Using Positional Embedding Interpolation (PE Interpolation)

➔ Adapt quickly to new resolutions with fewer training steps (1000 steps)

➔ Create high-resolution images without learning from scratch

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



(b)

(Skip)

PE Interpolation?

Previous limitation

If learned location embedding is 512×512 in size (LR Model),

Directly applying this embedding to a higher resolution (1024×1024) will result in mismatch

→ poor performance 😢

Apply PE Interpolation

Interpolate the existing position embedding to the 1024×1024 size.

This means that 512 values are **naturally converted to smooth values**

in the process of scaling to 1024.

Weak-to-Strong Training Strategy

Adapting model to KV compression

Using KV Token Compression

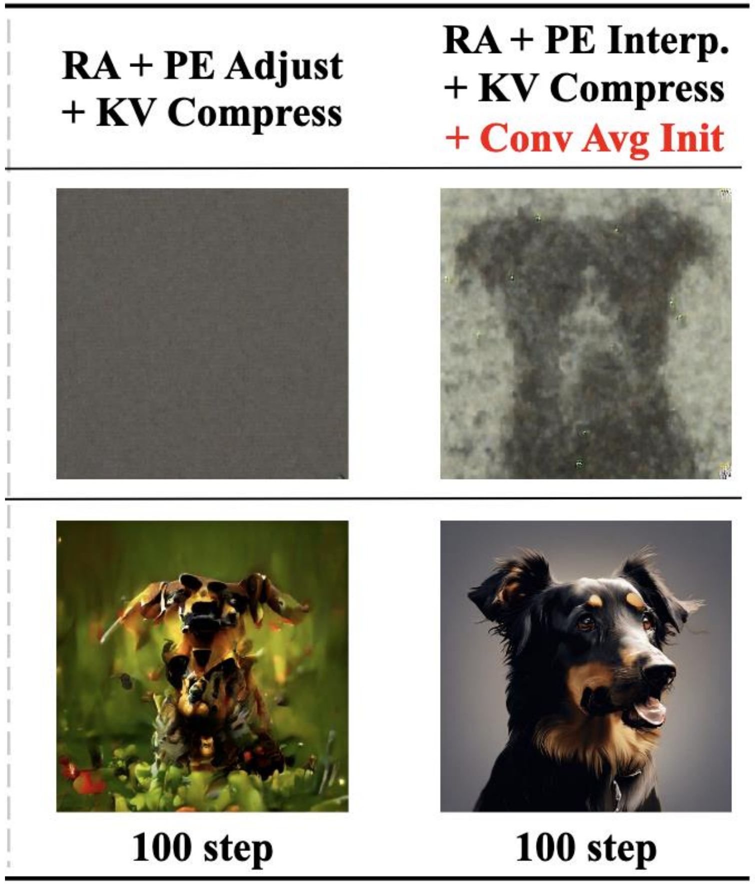
➔Risk of different structure 😢

➔Difficult to use the trained weights of PixArt-α

Using Conv Avg Init

➔ Set the weighting value to **$1/R^2$** to smooth the transition

➔ Preserving as much of the existing spatial information as possib



(c)

Experiment : Implementation Details

Model

Text-Incoder	Flan-T5-XXL (= PixArt-α)
VAE	Stable Diffusion XL(SDXL)
Base model	PixArt-α

Hardware

Training GPU (<=1K model)	32 NVIDIA Tesla V100
Training GPU (2K, 4K model)	16 NVIDIA A100
Optimization algorithms	CAME Optimizer

Evaluation Metrics

➔ 30,000 high quality dataset

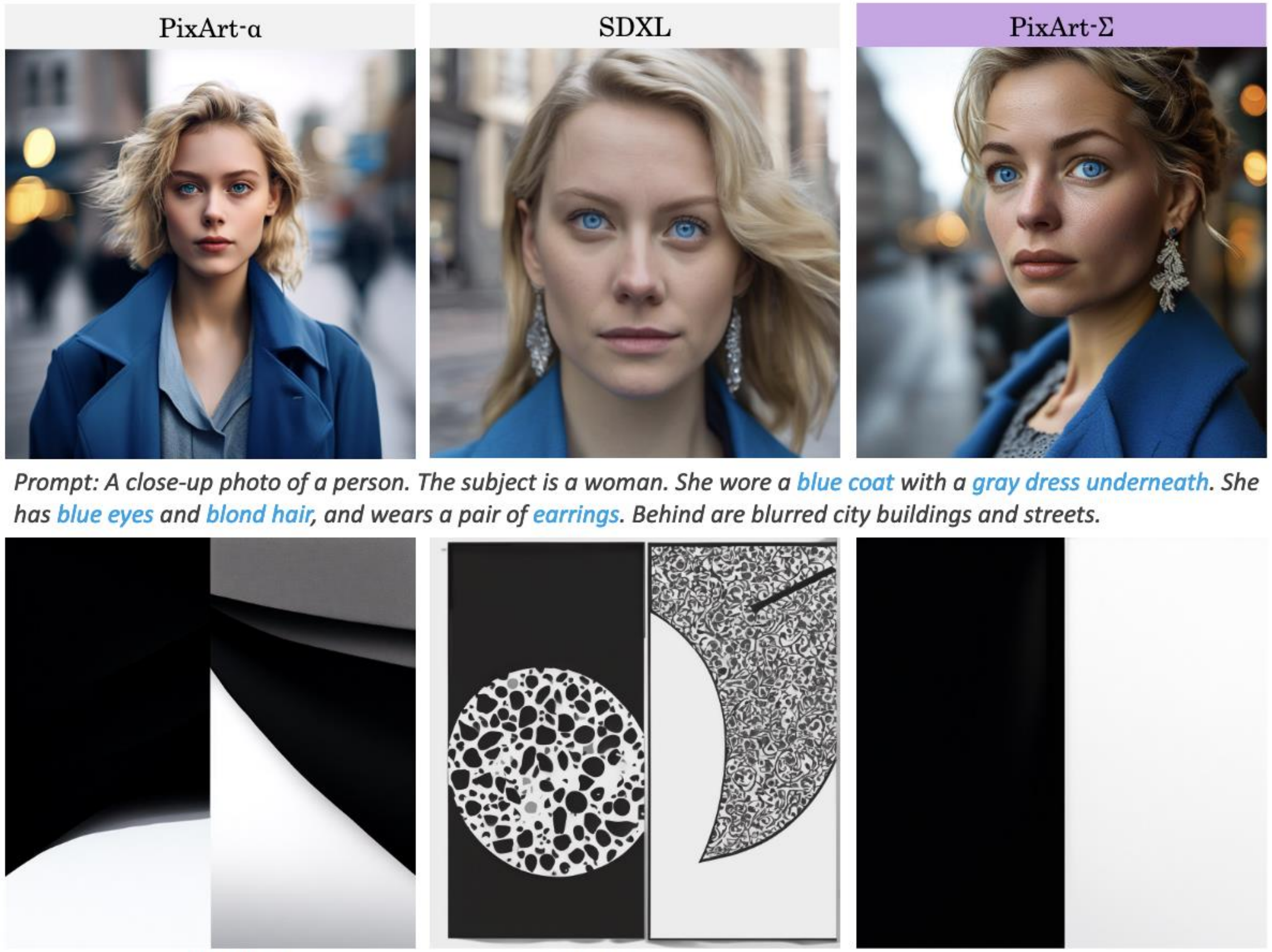
➔ benchmark the most powerful T2I models.

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

Experiment

Image Quality Assessment

Compared with open-source models



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

Experiment

Image Quality Assessment

competitive with these commercial products

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

Experiment

High-resolution Generation

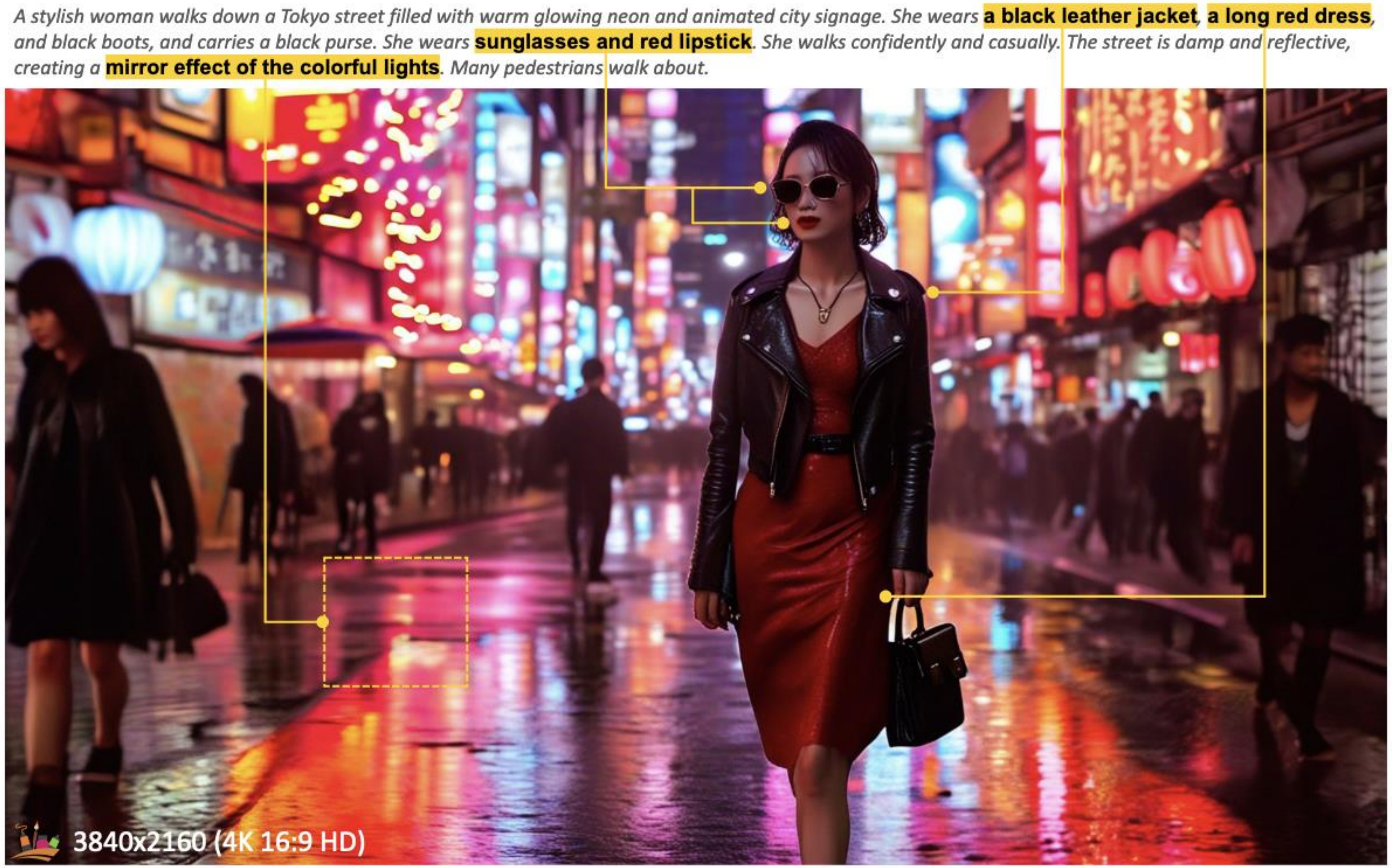


Fig. 2: 4K image generation with complex dense instructions. PixArt-Σ can directly generate 4K resolution images without post-processing, and accurately respond to the given prompt.

Experiment

Human/AI (GPT4V) Preference Study

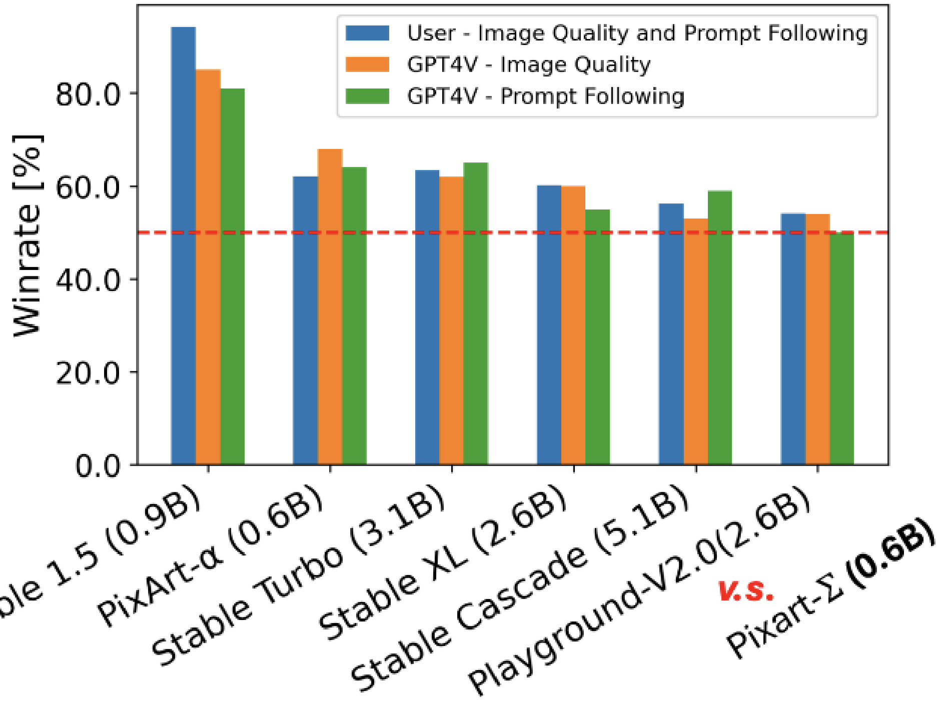
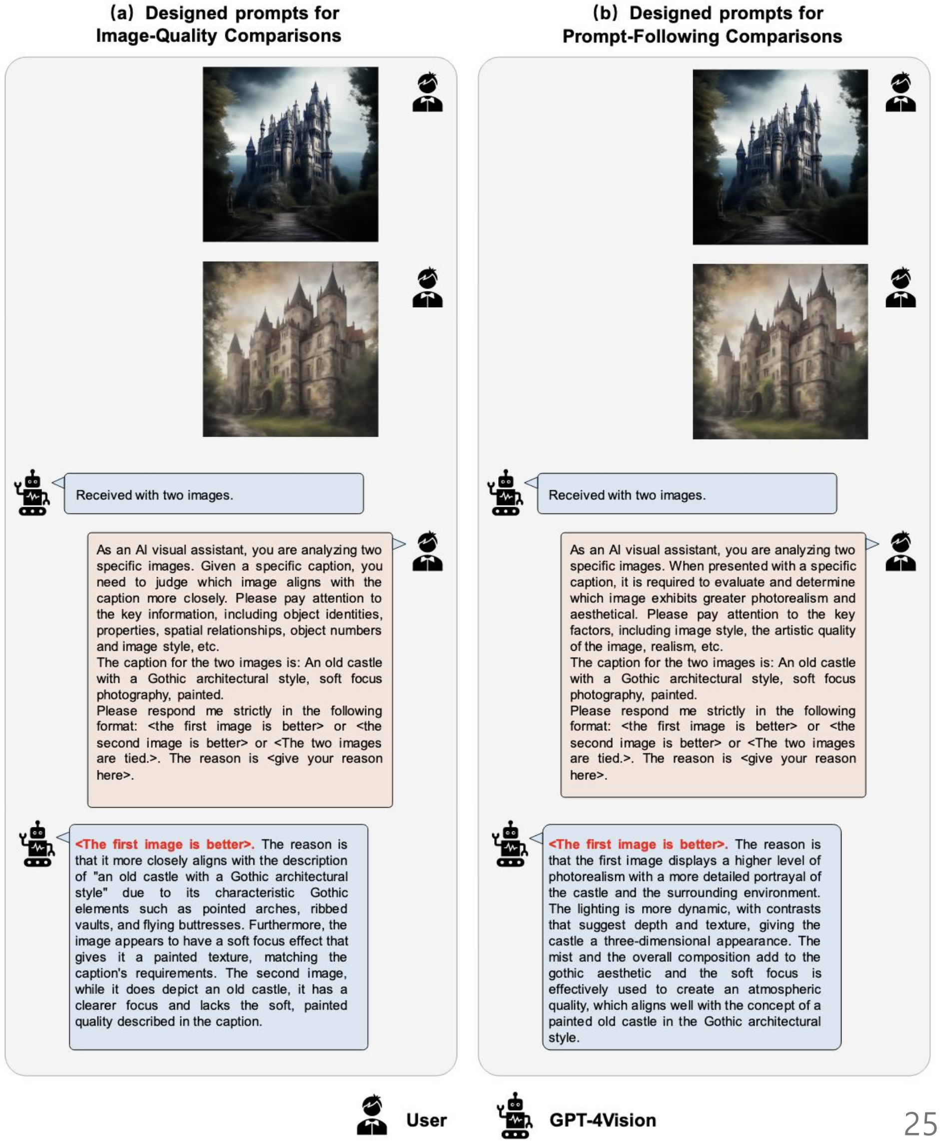


Fig. 9: Human(blue)/AI(orange and green) preference evaluation against currrent open T2I models. PixArt-Σ compares favorably against current state-of-the-art T2I models in both image quality and prompt-following.



Conclusion

Key : Weak-to-Strong Training + KV Token Compression

Category	PixArt-α	PixArt-Σ
Maximum Resolution	1K (1024×1024)	**4K (3840×2160) Supported**
Computation Optimization	X	**KV Token Compression** **(34% Reduction in Computation)**
VAE Model	Basic VAE	**SDXL VAE** **(Higher Quality Image Generation)**
Text Token Length	120 Tokens	**300 Tokens** **(More Precise Text-Image Alignment)**
Training Strategy	Standard Training	**Weak-to-Strong Training** **(Utilizing Pre-trained Model for Faster Training)**
Training Cost	High	**9% GPU Cost**

Limitations

Not on-device	◆Can't run on mobile and edge devices → Privacy concerns
Insufficient dataset	◆Using 33M data = less than Stable Diffusion v1.5 (2B data) → Quality degradation
speed issues	◆4K creation is possible but **not optimized for speed**

Future work

4K On-device diffusion

Privacy issue

- ◆ Handling photos is always a privacy risk
- ◆ There is no on-device 4K diffusion paper now
- ◆ Cloud can use your photo 😠

On-device

- ◆ Experimenting on-device with this model
- ◆ Identify issues on-device (latency, battery, memory etc.)

Optimization

- ◆ Optimized to work on **smartphones**
- ◆ Optimize by applying modern paper techniques like '*MobileDiffusion*'

vs MobileDiffusion

Category	PixArt-Σ	MobileDiffusion
Model Architecture	Diffusion Transformer (DiT) based	Latent Diffusion + Optimized UNet
Text Encoder	Flan-T5-XXL	CLIP-ViT/L14
Image Resolution	Direct 4K (3840×2160)	512×512
KV Token Compression	✅	❌
Model Size	0.6B	386M
VAE (Autoencoder)	SDXL VAE	Lightweight VAE
Resolution Upscaling Method	PE Interpolation	Fixed at 512px (No upscaling)
Computation Optimization	Weak-to-Strong Training (Reuses pre-trained models)	Transformer block removal + Convolution-based optimization
On-Device Execution	❌ Requires high-performance GPU	✅ iPhone 15 Pro, Samsung S24 etc.
Training Dataset Size	33M (Includes 4K)	150M
Image Quality Evaluation (FID Score)	8.23	11.67 (1-step) / 8.65 (50-step DDIM)
Text-Image Alignment (CLIP Score)	0.2797	0.320 (1-step) / 0.325 (50-step DDIM)
Generation Speed	Slow on high-end GPU for 4K	0.2s on iPhone 15 Pro

FID : Fréchet Inception Distance / CLIP (Contrastive Language-Image Pretraining Score)

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

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, Zhenguo Li

Jeonghoon Park, happypjh2001@unist.ac.kr
Undergraduate Research Intern,
Ubiquitous Artificial Intelligence Lab,
Department of Computer Science and Engineering,
Ulsan National Institute of Science and Technology