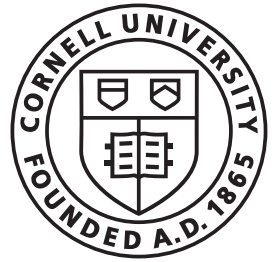
ICML
International Conference
On Machine Learning

ICL
Intelligent Comput.
imaging Lab.



Skrr: Skip and Re-use Text Encoder Layers for Memory Efficient Text-to-Image Generation

Hoigi Seo[1*], Wongi Jeong[1*], Jae-sun Seo[2] and Se Young Chun[1,3†]

* Authors contributed equally, † Corresponding author

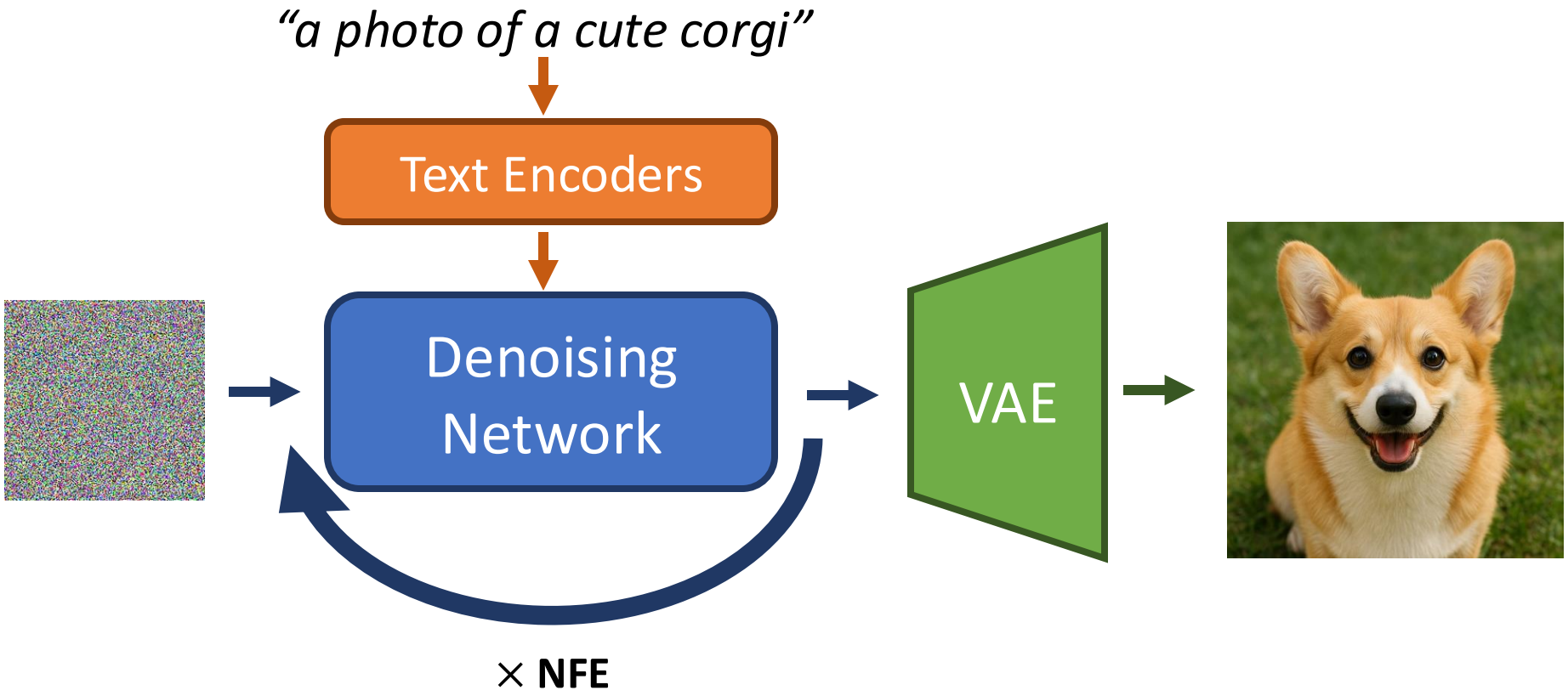
[1]Dept. Of Electrical and Computer Engineering , Seoul National University, Republic of Korea

[2]School of Electrical and Computer Engineering, Cornell Tech, USA

[3]INMC & IPAI , Seoul National University, Republic of Korea

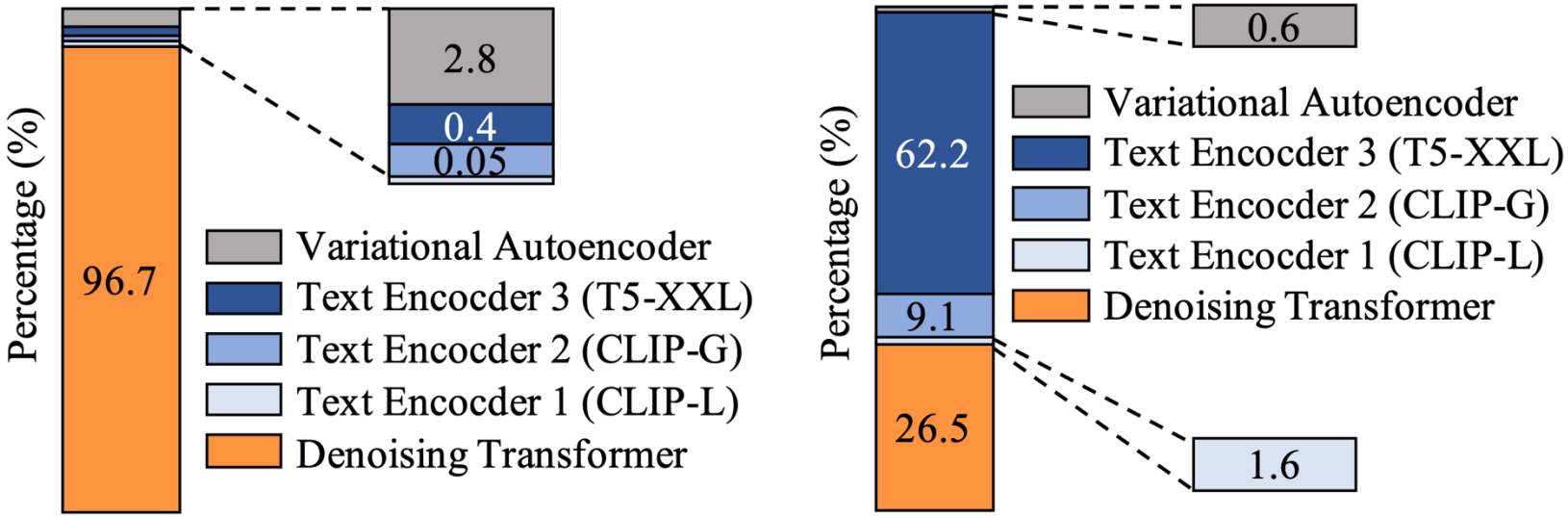
Text-to-Image (T2I) Diffusion Models

- Generates images aligned with given prompts
  - It usually consists of variational autoencoder (VAE), text encoders and denoising network.



Huge Text Encoders in T2I Diffusion Model

- Modern T2I diffusion models often have **huge** text encoders
  - Although the text encoder requires relatively less computation during image synthesis, it accounts for a large proportion of the total parameters in the pipeline.



(a) FLOPs ratio. (b) Parameter ratio.

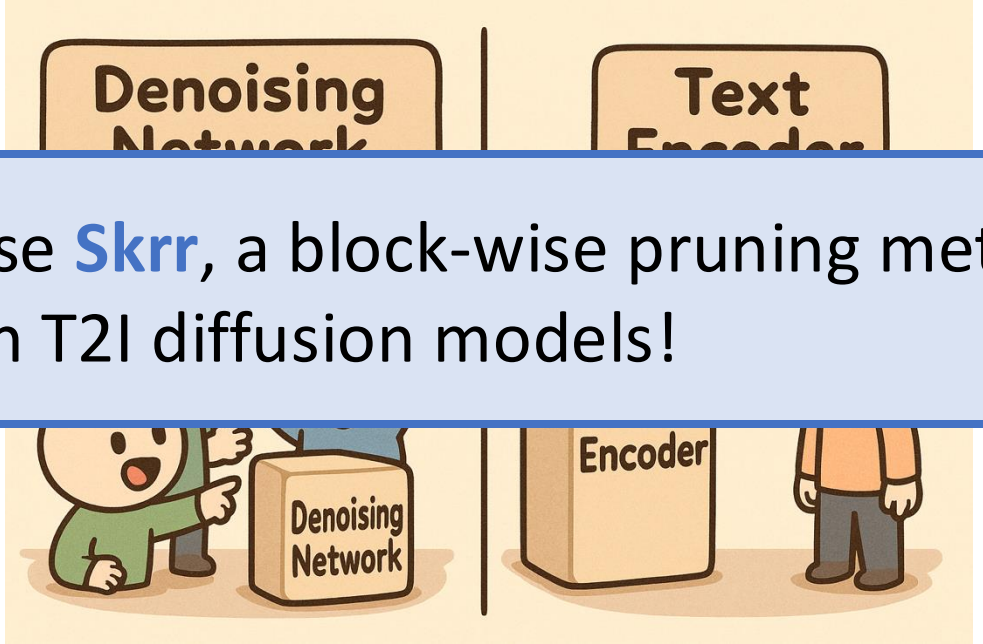
FLOPs and parameter ratio of Stable Diffusion 3 (SD3)

Esser, Patrick, et al. "Scaling rectified flow transformers for high-resolution image synthesis." ICML. 2024.

Compressing Text Encoder with Pruning

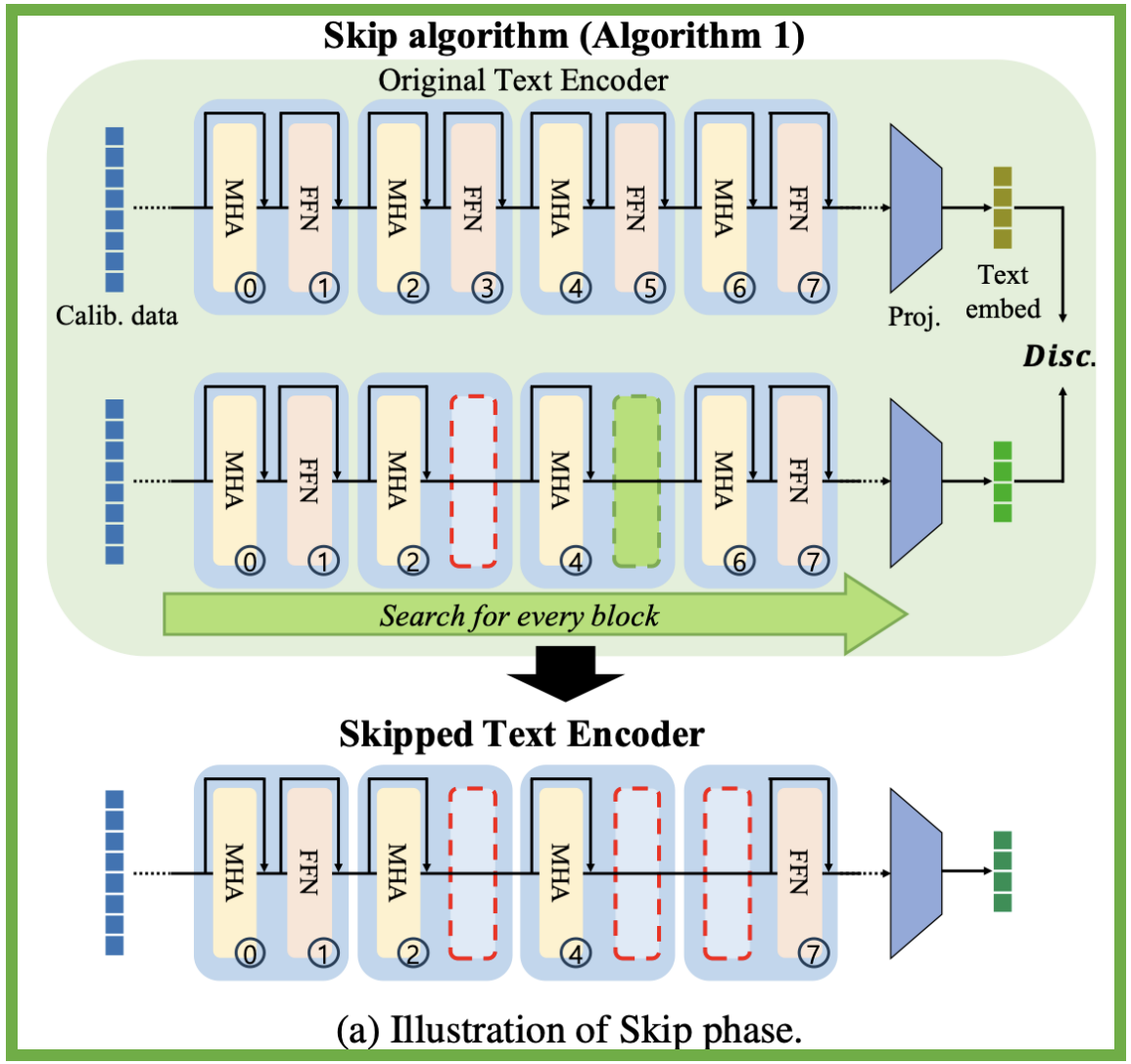
☐ Compressing text encoders in T2I diffusion models are under-explored for memory efficient image synthesis

- Most previous works on compressing with pruning T2I diffusion models focuses on compressing denoising models.
- …yet text encoders occupy the most memory despite their lower computational cost.

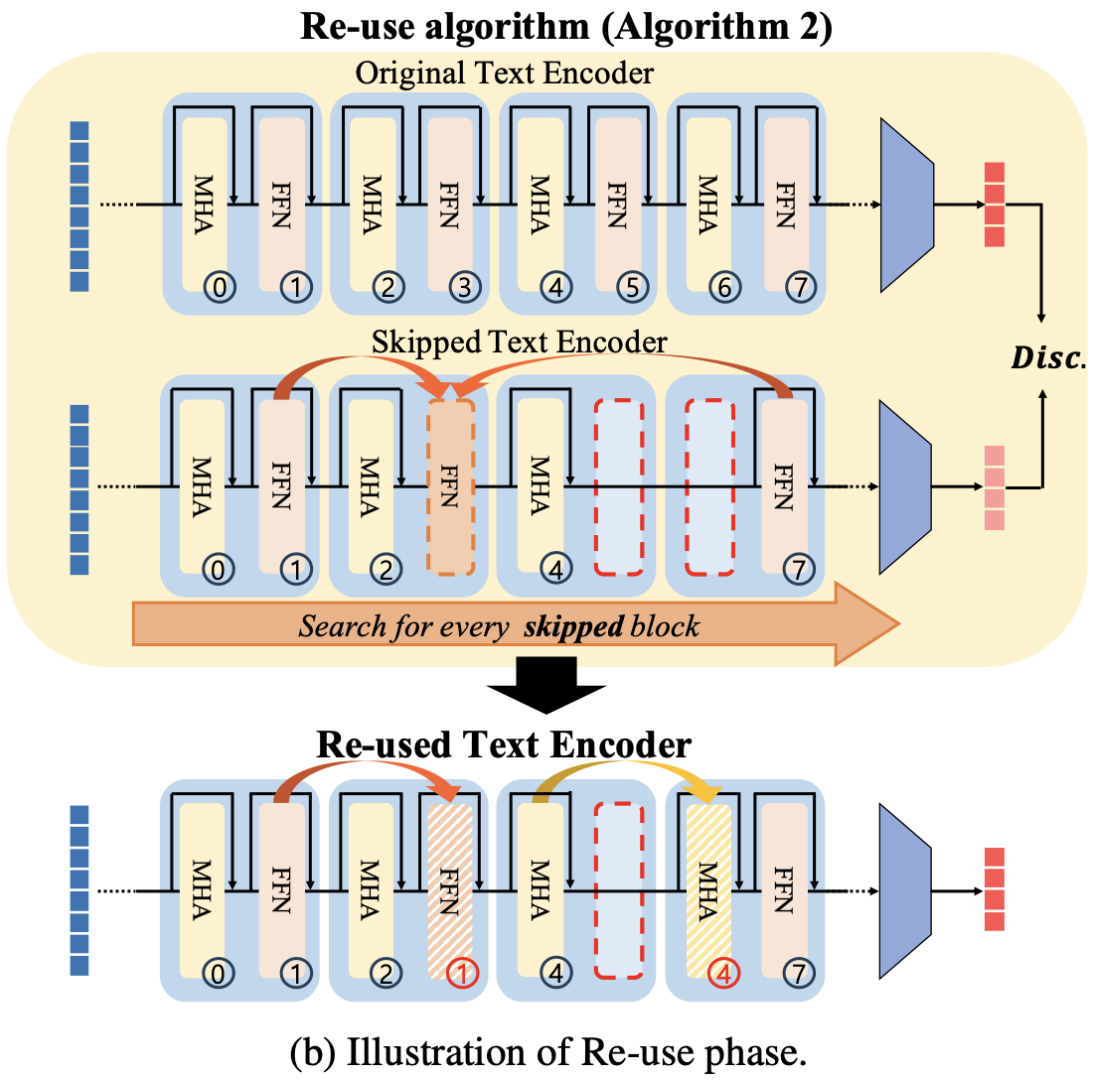


Here we propose **Skrr**, a block-wise pruning method tailored for text encoders in T2I diffusion models!

Overall pipeline of Skrr



(a) Illustration of Skip phase.



(b) Illustration of Re-use phase.

Discrepancy Metrics

☐ Discrepancy Metrics

- $f = \text{proj}(E(c; \theta_{\text{text}}); \theta_{\text{denoise}})$
- $\text{Metric}_1(f_{dense}, f_{skip})$: $1 - \frac{f_{dense} \cdot f_{skip}}{|f_{dense}||f_{skip}|}$
- $\text{Metric}_2(f_{dense}, f_{skip})$: $\text{MSE}(f_{dense}, f_{skip})$

(a) Original image.

(b) 7^{th}, 22^{th} removed.
Metric_1: 0.15
Metric_2: 0.002

(c) 3^{rd}, 5^{th} removed.
Metric_1: 0.11
Metric_2: 0.04

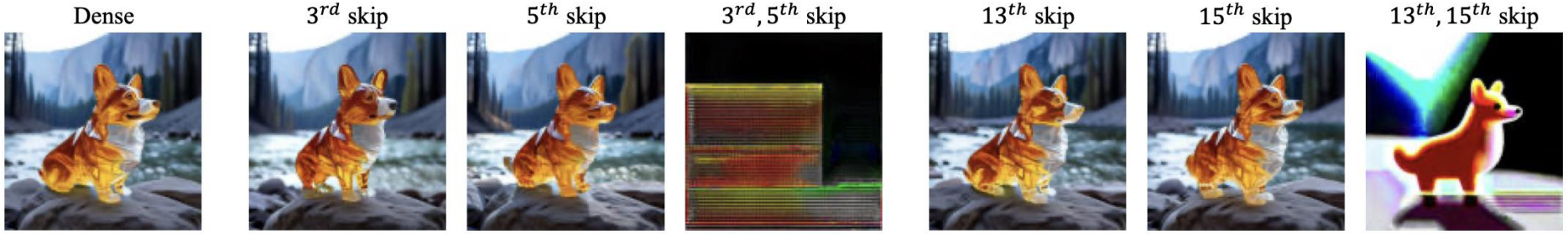
Block Interaction and Beam Search

☐ There are some blocks with interactions

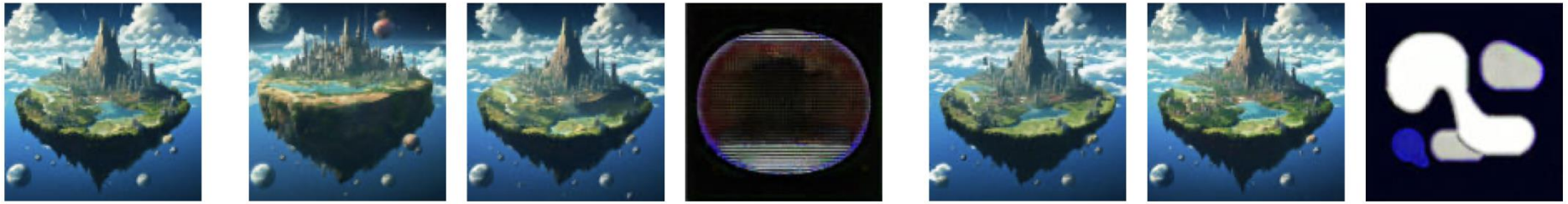
- Some blocks can be safely pruned individually, but issues arise when pruned together.

→It is necessary to consider multiple pruning paths simultaneously.

- We mitigated this by using a beam search-based algorithm.

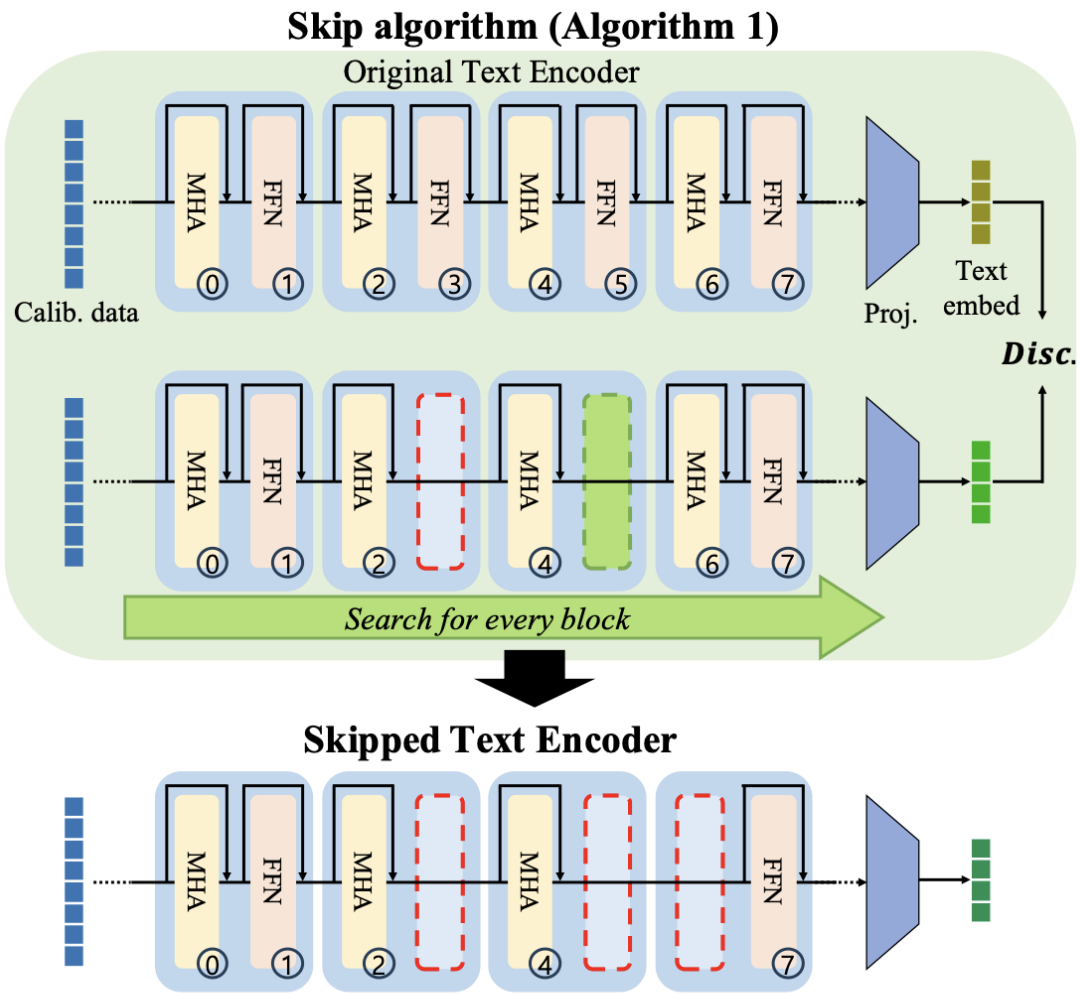


"Color photo of a corgi made of transparent glass, standing on the riverside in Yosemite National Park."

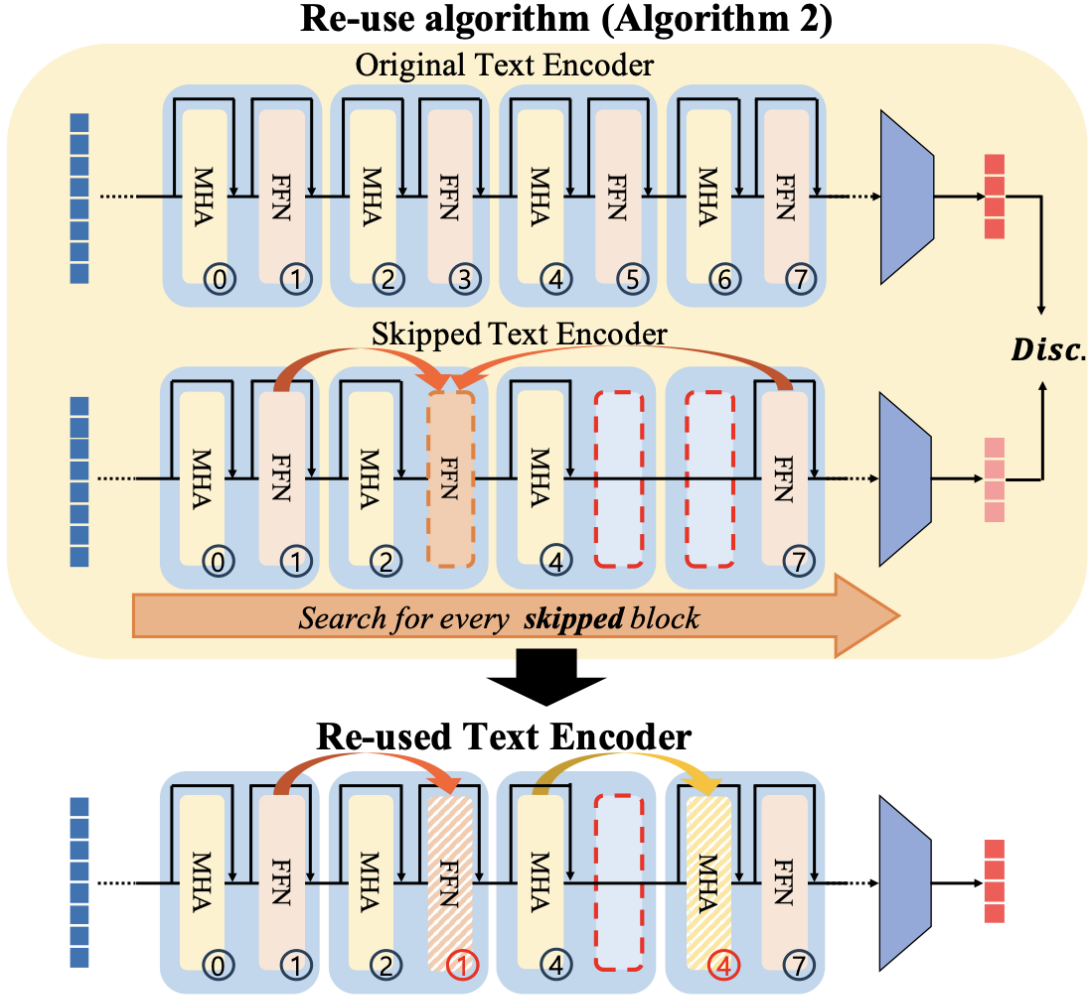


"Game-Art - An island with different geographical properties and multiple small cities floating in space."

Overall pipeline of Skrr



(a) Illustration of Skip phase.



(b) Illustration of Re-use phase.

Theoretical Validation of Re-use

☐ Error bound of two transformers

Lemma 3.1 *(Error bound of two transformers).* *Let* $\mathcal{M}: (x, \theta) \mapsto \mathbb{R}^d$ *be an L-block transformer with input* $x \in \mathbb{R}^d$ *and parameter set* $\theta = (\theta_1, \ldots, \theta_L)$ *defined as:*

$$\mathcal{M} = \left((F_L + I) \circ (F_{L-1} + I) \circ \cdots \circ (F_1 + 1)\right)$$

where $F_i: (z_i, \theta_i) \mapsto \mathbb{R}^d$ *is the* i*-th block with parameters* θ_i*, and* $z_i \in \mathbb{R}^d$*. Assume that* F_i *is* L_i*-Lipschitz in* z_i *and* M_i*-Lipschitz in* θ_i*, Then, for any two parameter sets* $\theta = (\theta_1, \ldots, \theta_L)$ *and* $\hat{\theta} = (\widehat{\theta_1}, \ldots, \widehat{\theta_L})$*, the following holds:*

$$\left|\left|\mathcal{M}(x;\theta) - \mathcal{M}(x;\hat{\theta})\right|\right| \leq \sum_{i=1}^{L} \left(\prod_{k=i+1}^{L} (1 + L_k) \right) M_i ||\theta_i - \widehat{\theta_i}||$$

Theoretical Validation of Re-use

☐ Tighter error bound of Re-use

Theorem 3.2 *(Tighter error bound of Re-use).* *Under the assumptions of Lemma 3.1, let be the* θ_i^* *parameters of the reused* F_i*. Define* U_{skip} *as the error bound for the compressed model with Skip alone and* $U_{skip,\,Re-use}$ *as the error bound for the compressed model with Skip and Re-use. If* $||\theta_i - \theta_i^*|| < ||\theta_i||$ *then the following holds:*

$$U_{skip,\,Re-use} < U_{skip}$$

Quantitative Results - Performance

☐ Experiments on PixArt-Σ

Method	Sparsity (%)	FID ↓	CLIP ↑	DreamSim ↑	GenEval ↑						
					Single	Two	Count.	Colors	Pos.	Color attr.	Overall
Dense	0.0	22.89	0.314	1.0	0.988	0.616	0.475	0.795	0.108	0.255	0.539
ShortGPT	24.3	24.96	0.309	0.753	0.944	0.381	0.431	0.715	0.033	0.083	0.431
	32.4	27.28	0.294	0.651	0.834	0.197	0.291	0.537	0.048	0.038	0.324
	40.5	55.26	0.215	0.357	0.306	0.025	0.090	0.100	0.0	0.0	0.087
LaCo	24.3	19.45	0.311	0.726	0.909	0.336	0.394	0.713	0.065	0.128	0.424
	32.4	24.70	0.303	0.677	0.781	0.227	0.250	0.606	0.043	0.040	0.325
	40.5	21.60	0.291	0.620	0.784	0.162	0.150	0.489	0.030	0.033	0.275
FinerCut	26.3	20.66	0.313	0.798	0.947	0.465	0.394	0.737	0.103	0.105	0.458
	32.2	20.49	0.313	0.771	0.903	0.409	0.344	0.697	0.078	0.128	0.426
	41.7	20.36	0.308	0.731	0.841	0.306	0.306	0.628	0.050	0.073	0.367
Skrr (Ours)	27.0	20.15	0.315	0.800	0.956	0.434	0.425	0.763	0.095	0.145	0.471
	32.4	20.19	0.313	0.775	0.928	0.397	0.413	0.774	0.100	0.118	0.455
	41.9	19.93	0.312	0.741	0.913	0.410	0.450	0.755	0.055	0.068	0.442

Men, Xin, et al. "Shortgpt: Layers in large language models are more redundant than you expect." arXiv preprint arXiv:2403.03853 (2024).
Yang, Yifei, Zouying Cao, and Hai Zhao. "Laco: Large language model pruning via layer collapse." arXiv preprint arXiv:2402.11187 (2024).
Zhang, Yang, et al. "Finercut: Finer-grained interpretable layer pruning for large language models." arXiv preprint arXiv:2405.18218 (2024).

Quantitative Results - Efficiency

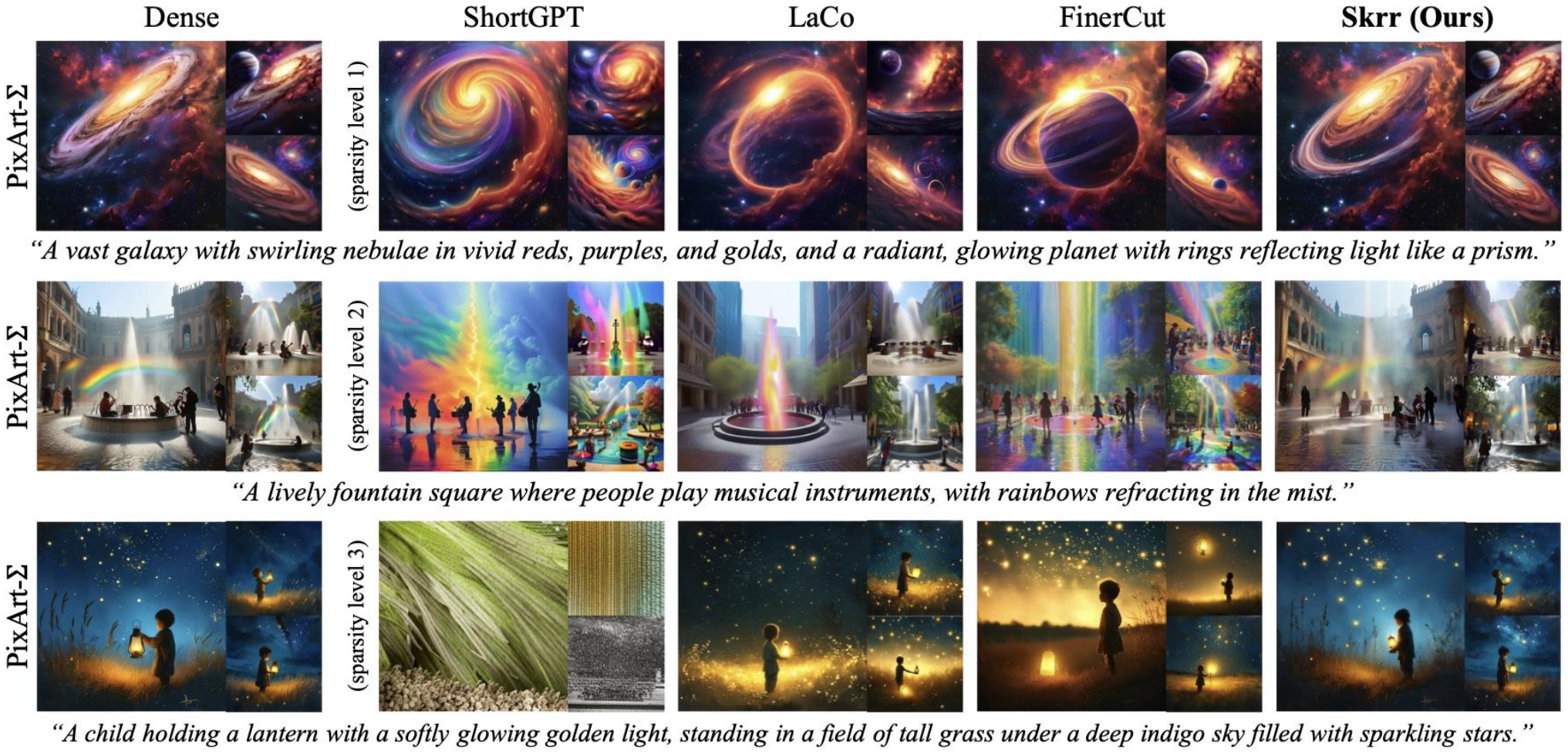
☐ Experiments on PixArt-Σ

- Similar sparsity, number of parameters and memory usage.
- Slightly increased FLOPs due to Re-use phase

Method	Sparsity (%)	Param. (B)	Mem. (GB)	TFLOPs
Dense	0.0	5.42	10.18	91.94
ShortGPT	40.5	3.49	6.59	91.79
FinerCut	41.7	3.43	6.48	91.74
Skrr (Ours)	41.9	3.43	6.46	91.90

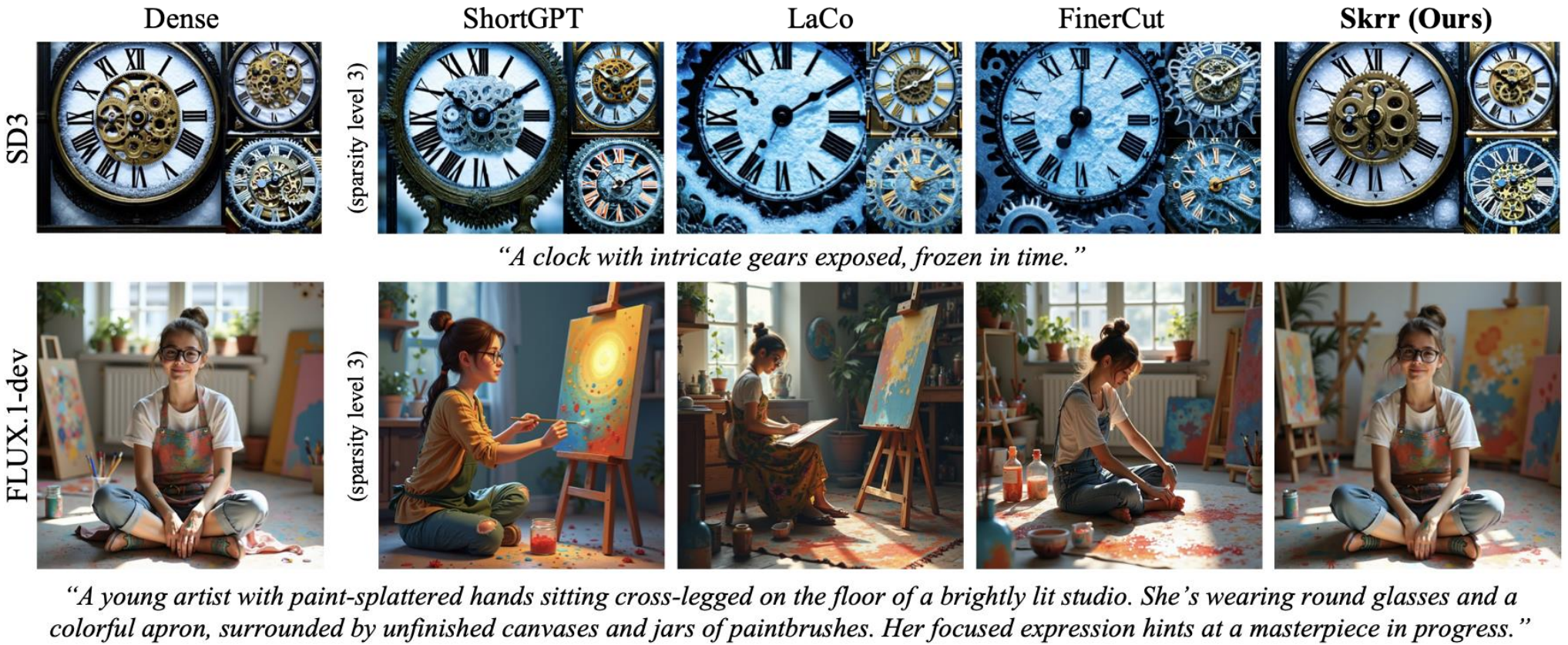
Qualitative Results

☐ Experiments with various sparsity



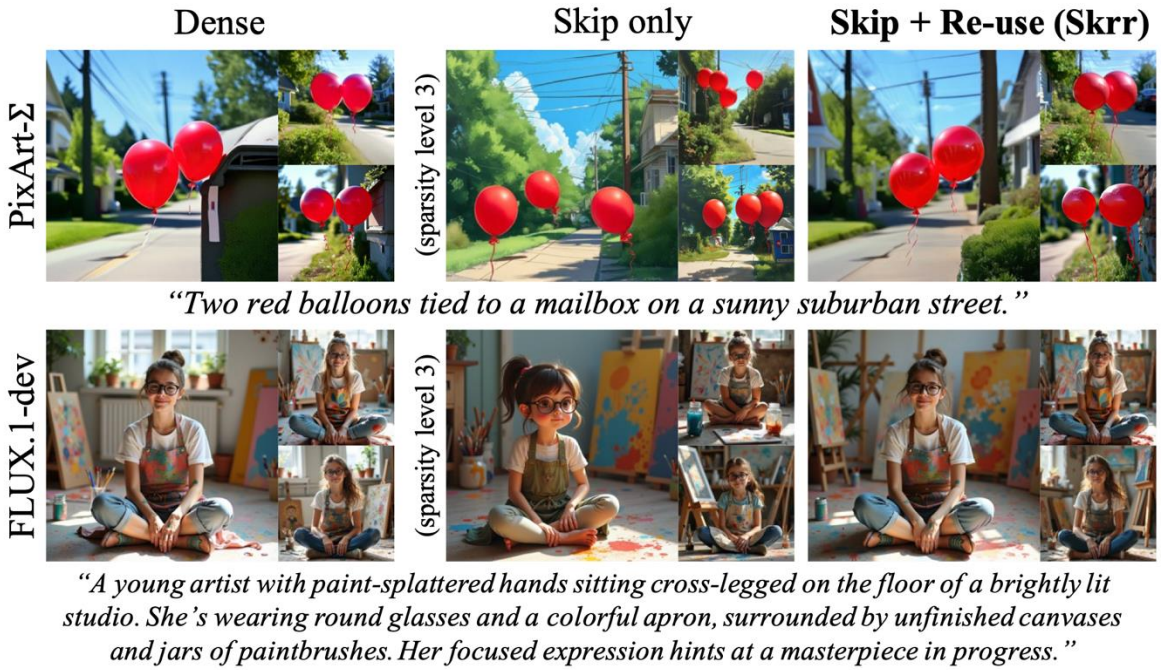
Qualitative Results

☐ Experiments with various baseline models (SD3, FLUX.1-dev)



Ablation Study

☐ Ablation study on Re-use and the size of beam k.



Effect of Re-use

Sparsity (%)	k	CLIP	DreamSim	GenEval		
				Single.	Count.	Colors.
41.9	1	0.310	0.737	0.900	0.372	0.739
41.9	2	0.312	0.757	0.912	0.450	0.731
41.9	3	0.313	0.746	0.912	0.450	0.755
40.7	4	0.310	0.739	0.925	0.328	0.707

Effect of the size of beam k

ICML
International Conference
On Machine Learning

Thank You!

ICL
Intelligent Comput. imaging Lab.

Skrr: Skip and Re-use Text Encoder Layers for Memory Efficient Text-to-Image Generation

For the paper and more results, please visit our project page! - ignoww.github.io/Skrr_project




Hoigi Seo[1*] **Wongi Jeong[1*]** **Jae-sun Seo[2]** **Se Young Chun[1,3†]**

Authors contributed equally, † Corresponding author

1 Dept. Of Electrical and Computer Engineering, Seoul National University, Republic of Korea
2 School of Electrical and Computer Engineering, Cornell Tech, USA
3 INMC & IPAI, Seoul National University, Republic of Korea

Acknowledgements

This work was supported in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2022M3C1A309202211). Also, the authors acknowledged the financial support from the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University.