Towards Understanding the Working Mechanism of Text-to-Image Diffusion Model

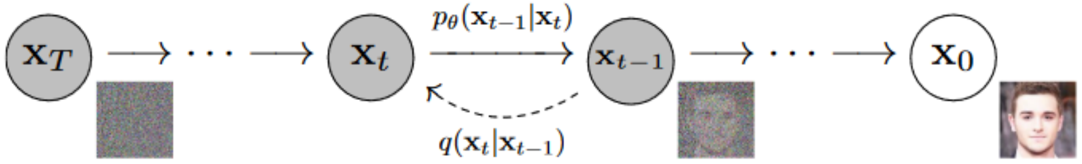
Mingyang Yi

Huawei Noah’ s Ark Lab

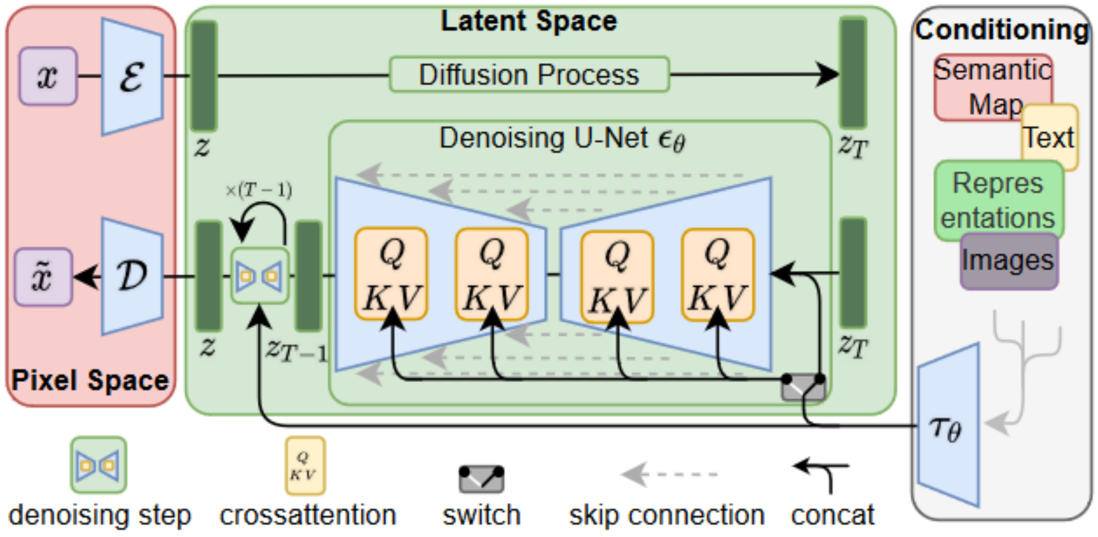
Joint work with Aoxue Li, Yi Xin, Zhenguo Li

Conditional Diffusion for T2I Generation

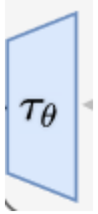
- The Process of Diffusion Model



- Condition Diffusion Model for Text-to-Image Generation



Cross-Attention



Text Encoder e.g., CLIP

How does T2I generation diffusion model works in practice?

Quickly Appeared Shape

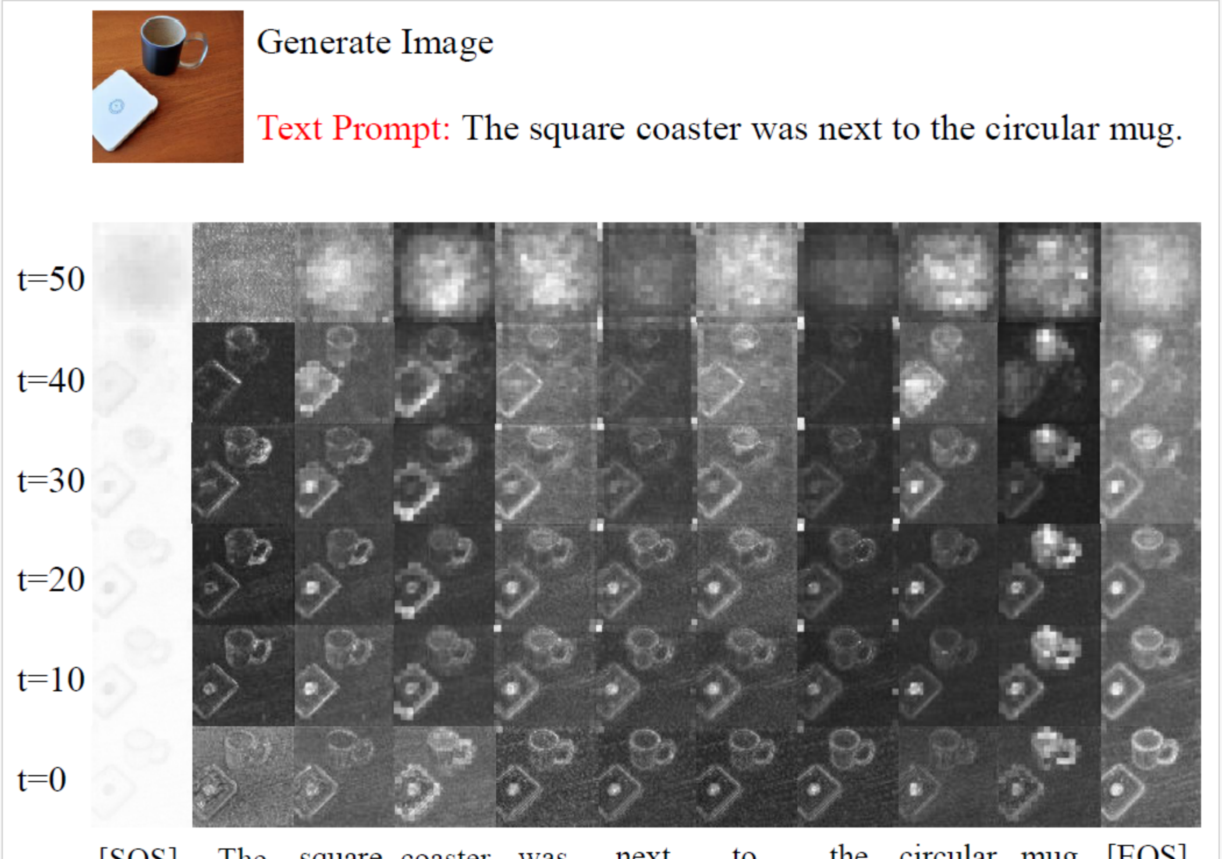
- Cross-Attention is Weighted Sum over Tokens

$$Q = W_Q \phi(x_t); K = W_K \mathcal{C}; V = W_V \mathcal{C}.$$

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^\top / \sqrt{d})V$$

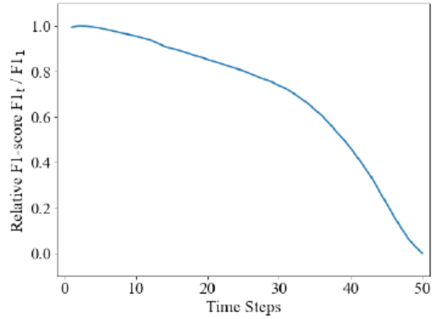
Cross-Attention Map: Image-Token Correlation

$\phi(x_t)$: Pixel **$\mathcal{C}$: Textual Prompt (Embedding)**



- The Shape is Quickly Recovered

The shape of image has been decided in the first few diffusion steps.



(b) Convergence of Cross-Attention Map

A Frequency Explain

high-freq -> shape low-freq -> details

- Noisy data and its Frequency

frequency

$$F_{\boldsymbol{x}_t}(u,v) = \frac{1}{MN}\sum_{k=0}^{M-1}\sum_{l=0}^{N-1} \boldsymbol{x}_t^{kl} \exp\left(-2\pi \mathrm{i}\left(\frac{ku}{M}+\frac{lv}{N}\right)\right)$$
$$= \sqrt{\bar{\alpha}_t} F_{\boldsymbol{x}_0}(u,v) + \sqrt{1-\bar{\alpha}_t} F_{\boldsymbol{\epsilon}_t}(u,v),$$

noisy data $\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t,$

- Energy of High-Freq v.s. Low-Freq

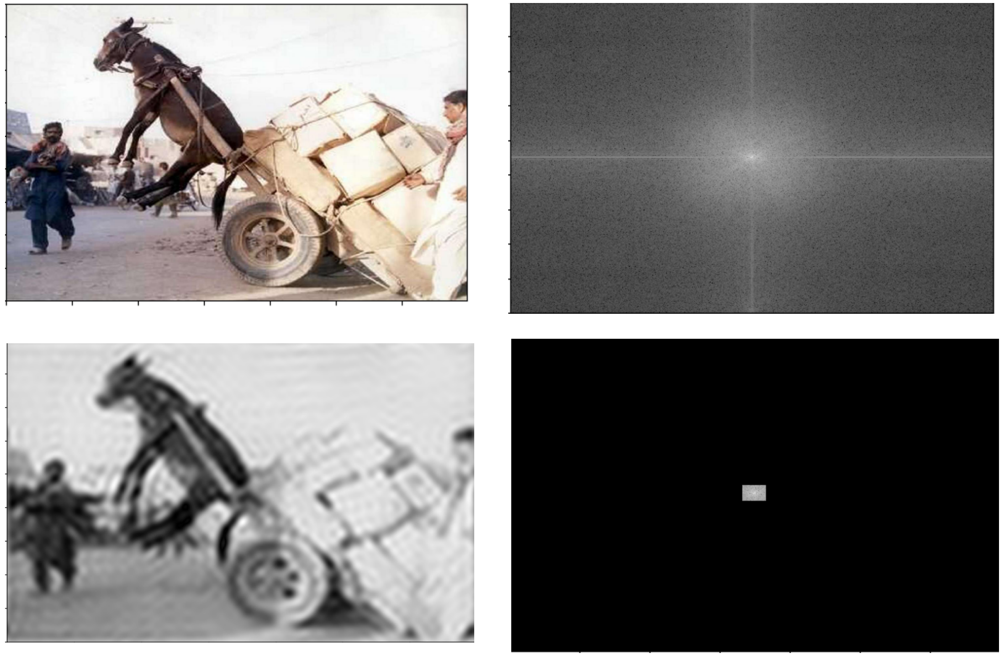
Proposition 1. *For all $u \in [M], v \in [N]$, with high probability, we have*

$$\|F_{\boldsymbol{\epsilon}_t}(u,v)\|^2 \approx \mathcal{O}\left(\frac{1}{\sqrt{MN}}\right).$$

white noise has more energy on high-freq.

natural noise has more energy on low-freq.

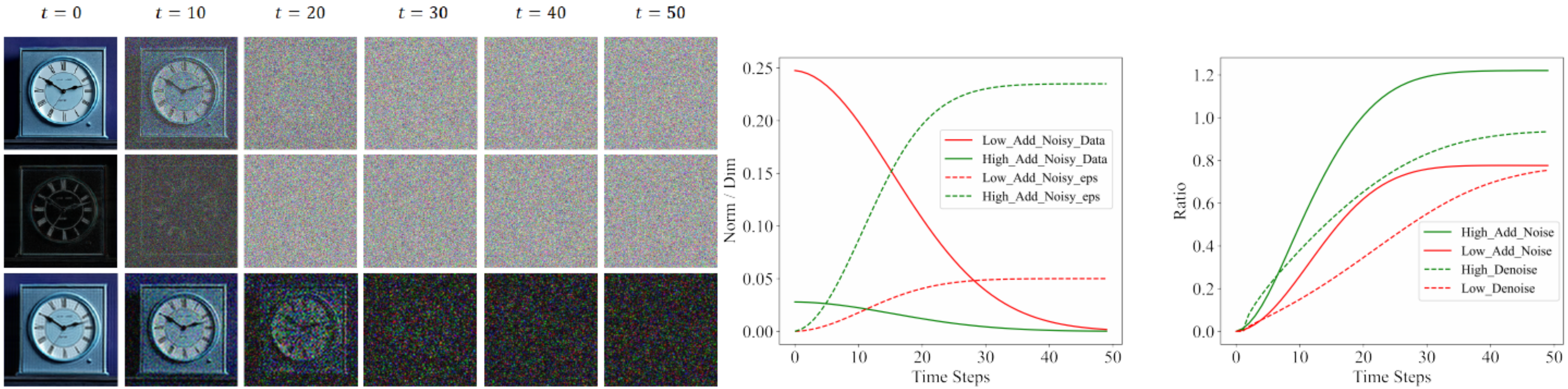
80% spectrum are high-freq



Low-Freq part of Image

Why First Shape then Details

- The high-freq part is quickly destroyed and will not be recovered until the end of reverse diffusion process. (vice-versa for low-freq)



(a) Noisy data and its high, low frequency parts (b) Norm of features $\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0$ and $\sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t$ (c) Ratio of high / low frequency parts variation

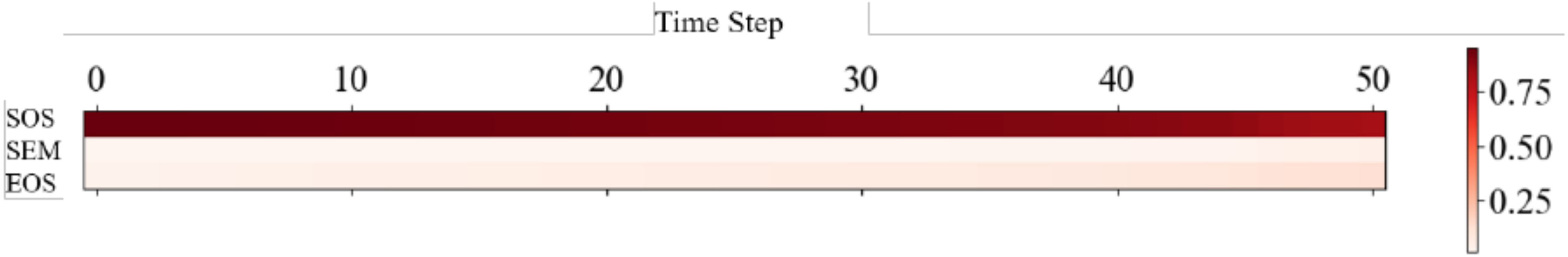
Focusing Shape and Details at beginning and end of diffusion, respectively.

Text Prompts Related to the Phenomenon

- Three Classes of Tokens

  Prompt: [SOS] a white vase [EOS] → [SOS] + Sem + [EOS]

- Auto-regressive encoder makes [SOS] contains no information



Weights on tokens, [SOS] adjust weights on cross-attention map.

[EOS] Decides Generation

- Generation Under Switched [EOS]

Prompt A + $[\text{EOS}]_A$ Prompt B + $[\text{EOS}]_B$ Prompt A + $[\text{EOS}]_B$ Prompt B + $[\text{EOS}]_A$

Prompt A: A blue bird
Prompt B: A brown chair

Prompt A: The sharp, angular edges of the city skyline pierced the clouds, a symbol of human innovation and progress.
Prompt B: The delicate, fluttering wings of the butterfly signaled the arrival of spring, a natural symbol of rebirth and renewal.

Observation I: [EOS] decides the overall T2I generation

Observation II: Slighter information in SEM is conveyed.

Table 1: The alignment of generated image with its source and target prompts. The prompts are constructed with switched [EOS].

Alignment \ Prompt	Source	Target
Text-CLIPScore ↑	0.2363	**0.2758**
BLIP-VQA↑	0.3325	**0.4441**
MiniGPT-CoT↑	0.6473	**0.7213**

Pay More Attention on [EOS]

When Does [EOS] Works

- **The [EOS] works on the first shape reconstruction stage.**

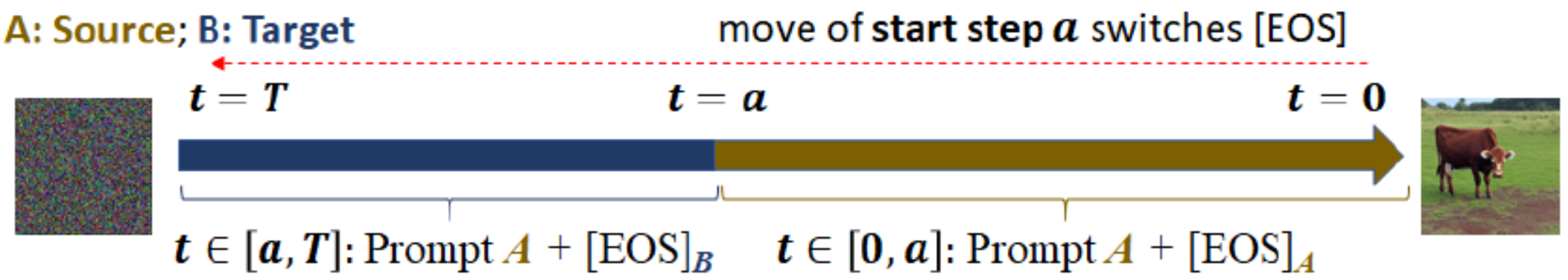
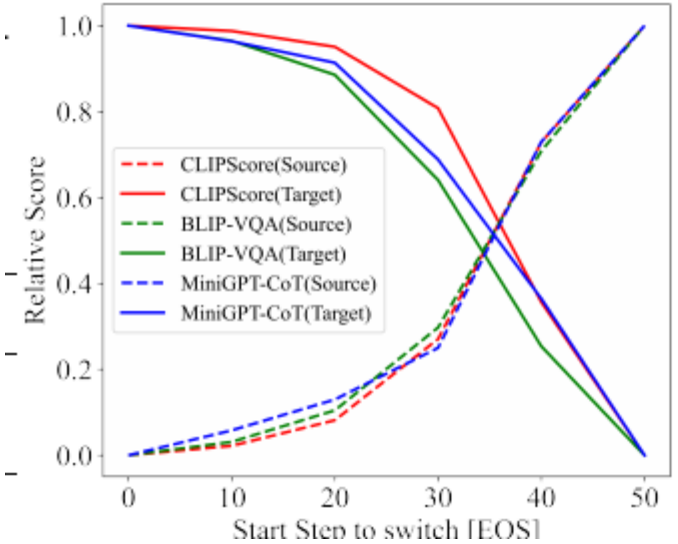


Figure 5: Desnoising process under text prompt with switched [EOS] in $[a, 50]$.



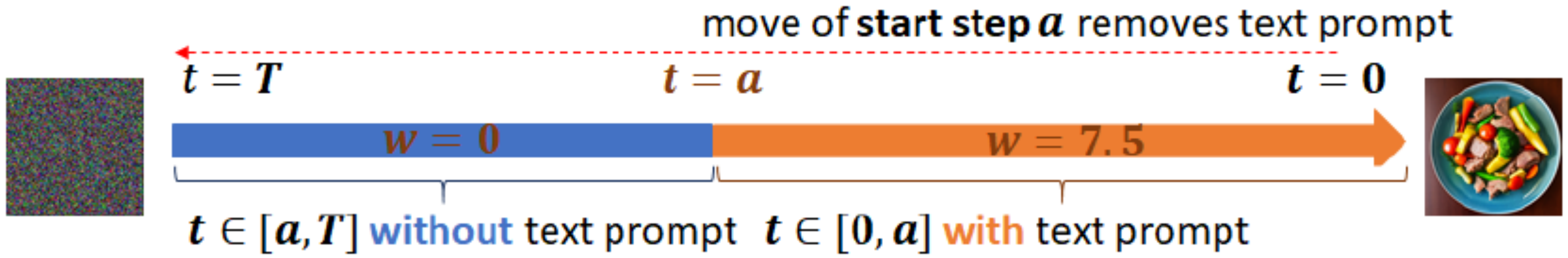
The effect of [EOS] is not disappeared until the removing it at the beginning of denoising process.

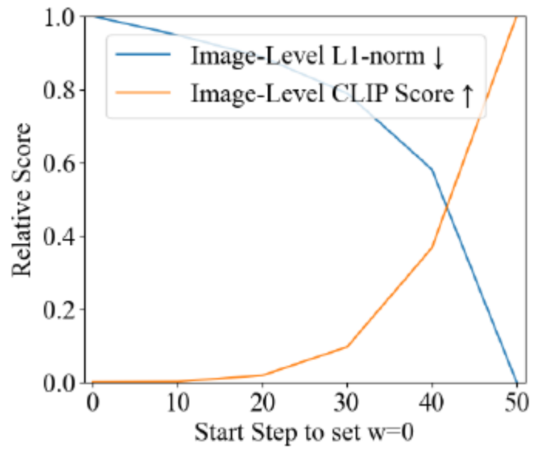
Text Information is Quickly Conveyed

- Noise Prediction

$$\epsilon_{\theta}(t, \boldsymbol{x}_t, \mathcal{C}, \emptyset) = \epsilon_{\theta}(t, \boldsymbol{x}_t, \emptyset) + w\left(\epsilon_{\theta}(t, \boldsymbol{x}_t, \mathcal{C}) - \epsilon_{\theta}(t, \boldsymbol{x}_t, \emptyset)\right),$$

- **Text Prompt Working on the Shape Reconstruction Stage**





Textual prompt is useless adding it at the beginning of diffusion process

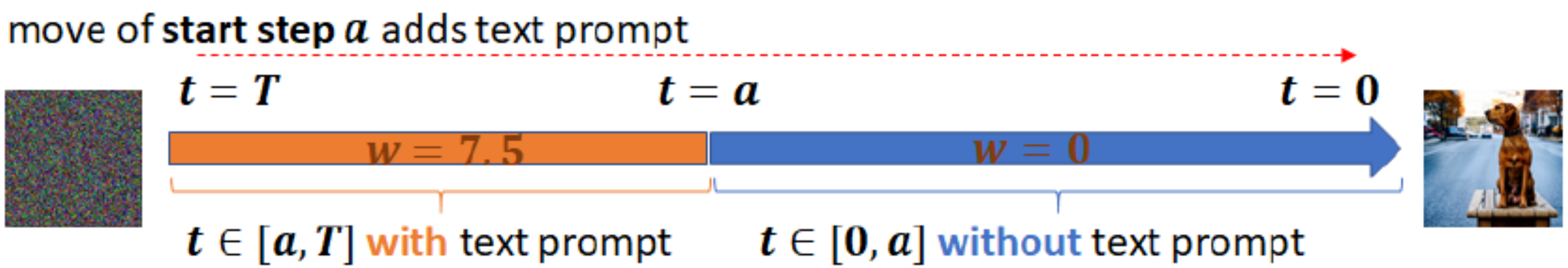
Conclusions

- The T2I Generation “First Overall Shape then Details” .
- The [EOS] Has More Impact.
- The Mainly Text Prompt Works in the First Stage.

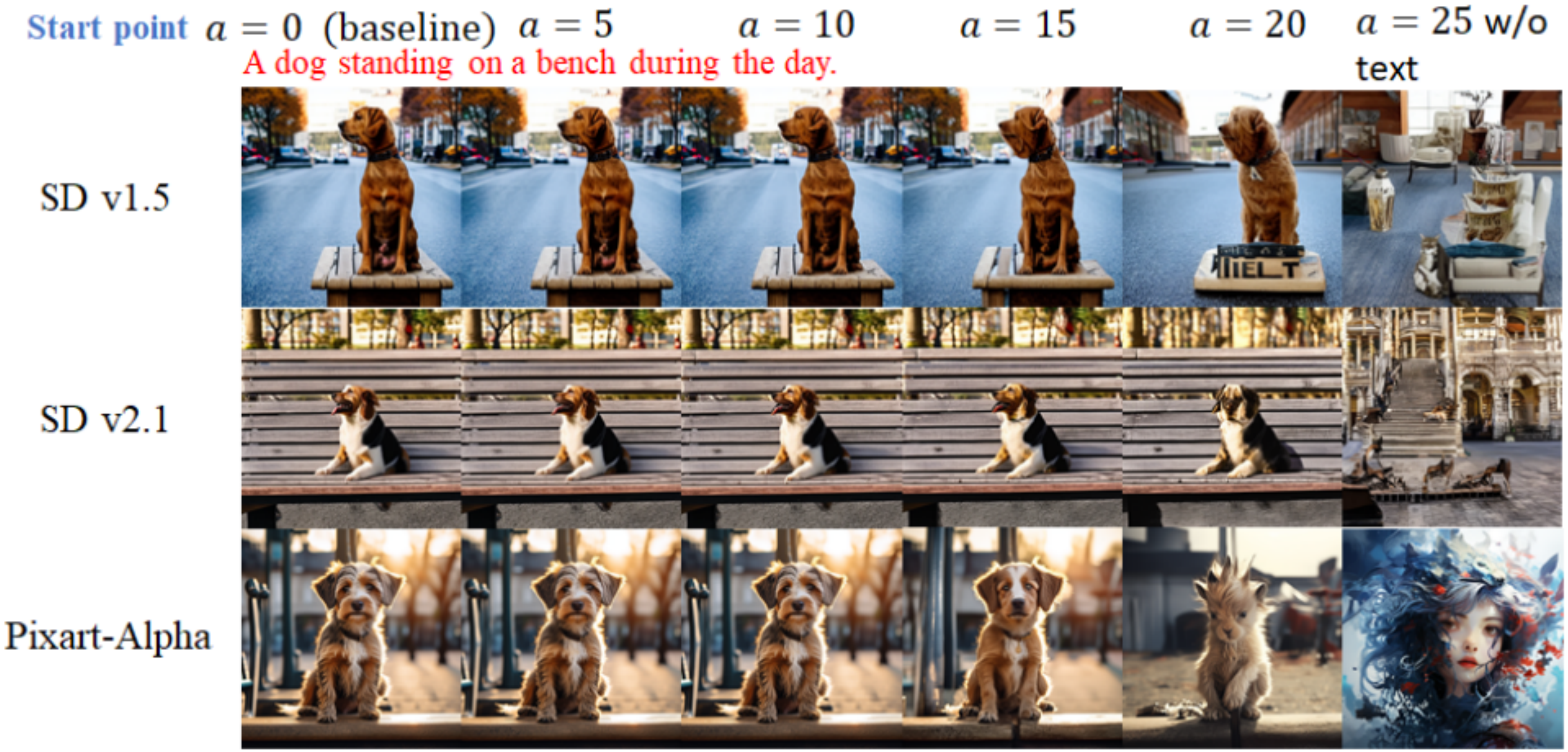
Application

- Accelerating Sampling with Removing Text Information

$$
\epsilon_{\theta}(t, \boldsymbol{x}_t, \mathcal{C}, \emptyset) = \begin{cases} \epsilon_{\theta}(t, \boldsymbol{x}_t, \emptyset) + w\left(\epsilon_{\theta}(t, \boldsymbol{x}_t, \mathcal{C}) - \epsilon_{\theta}(t, \boldsymbol{x}_t, \emptyset)\right) & a \le t; \\ \epsilon_{\theta}(t, \boldsymbol{x}_t, \emptyset) & 0 \le t < a. \end{cases}
$$



Results


Start point
$a = 0$ (baseline)
$a = 5$
$a = 10$
$a = 15$
$a = 20$
$a = 25$ w/o text
A dog standing on a bench during the day.
SD v1.5
SD v2.1
Pixart-Alpha

Thanks!