Attention in steps

Early stage : strongly relies on the text prompt

Later stage : text conditioning is almost entirely ignored, focus of high visual fidelity



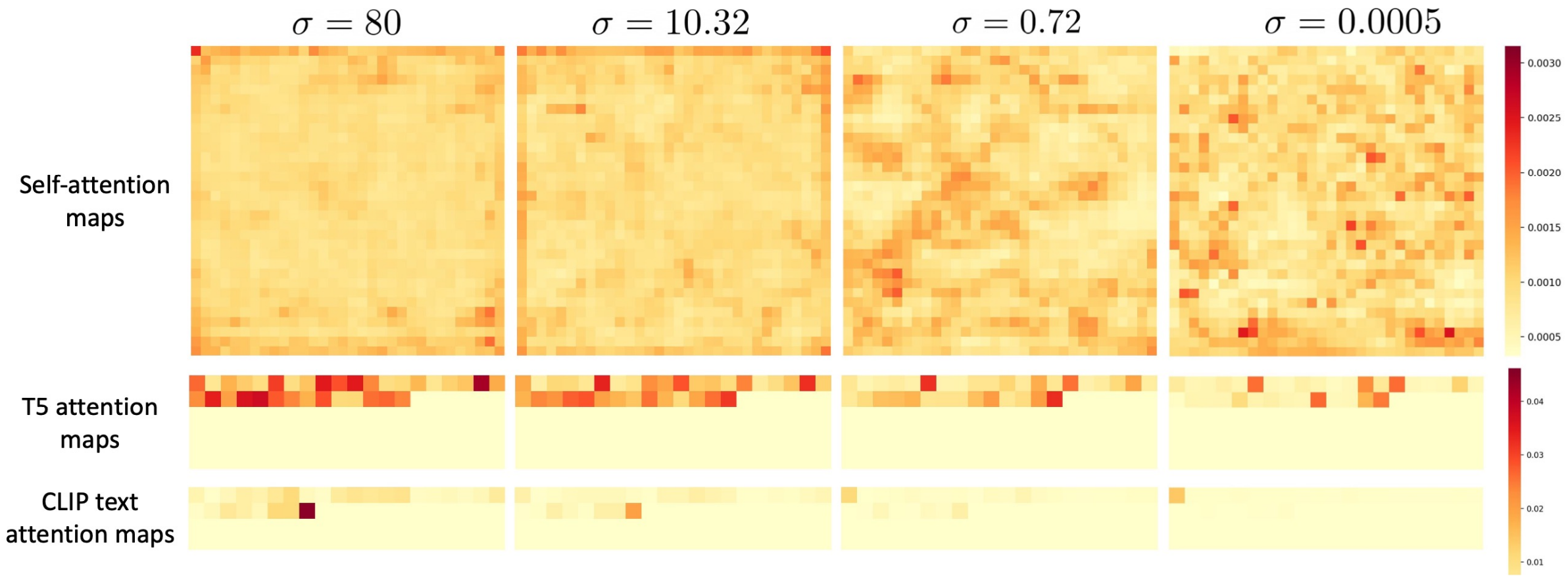


100%	93%	7%	70%	30%	40%	60%	100%

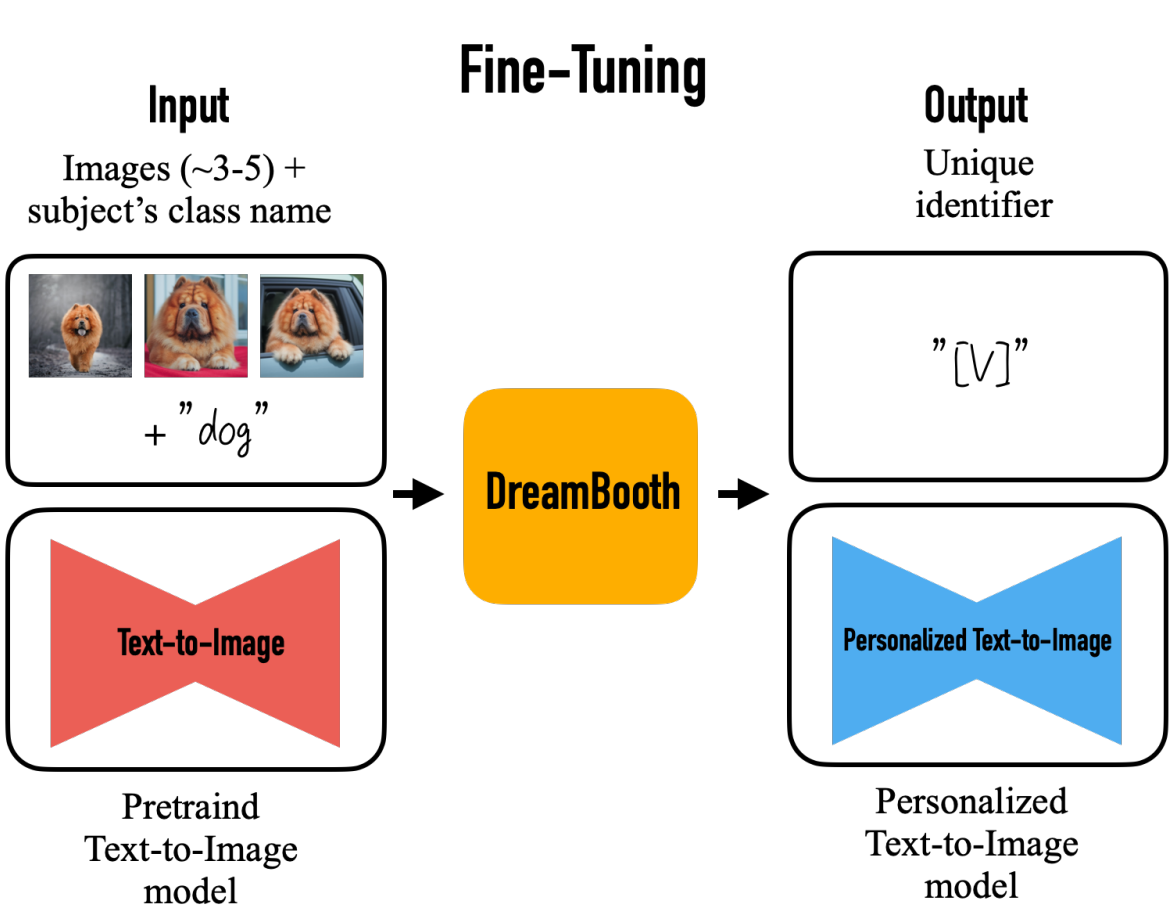
Prompt #1: *A 4k dslr image of a lemur wearing a red magician hat and a blue coat performing magic tricks with cards in a garden.*

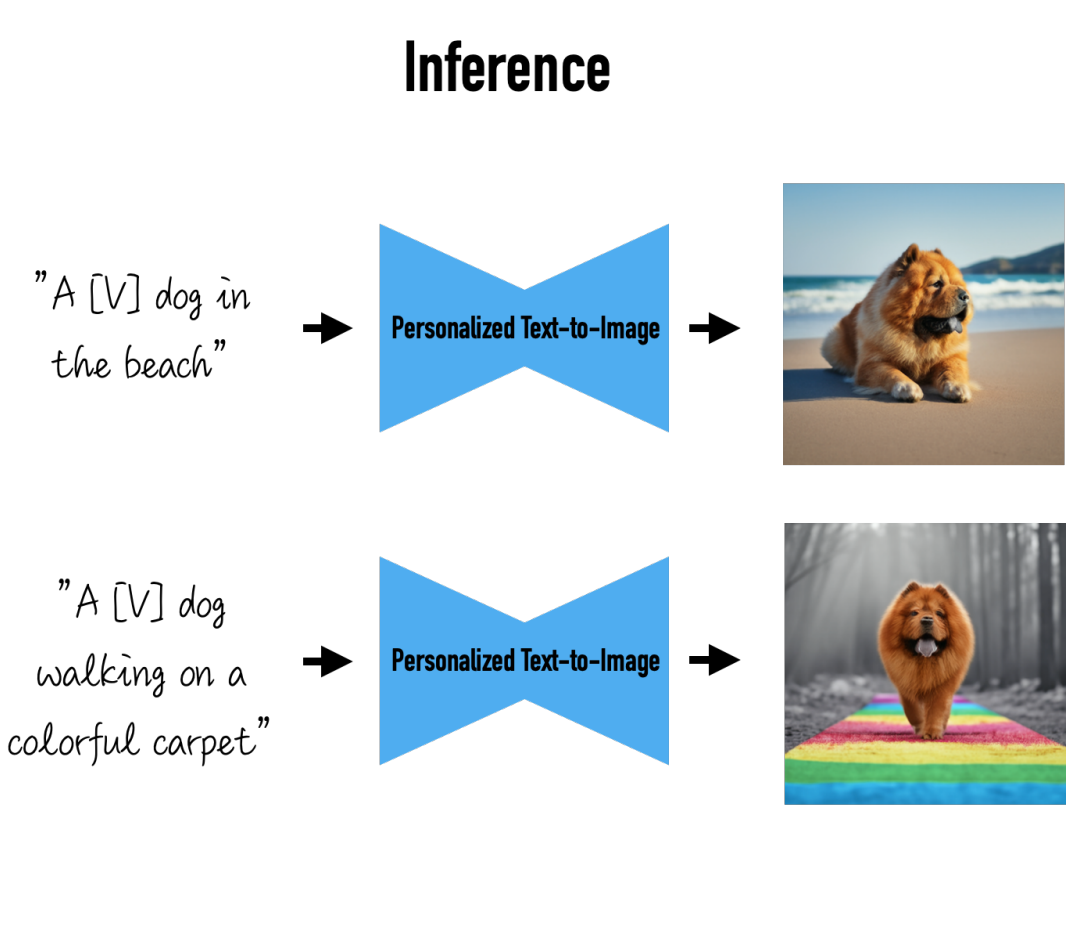
Prompt #2: *A 4k dslr photo of a cute lion cub floating in a bowl of honey.*

Attention in steps

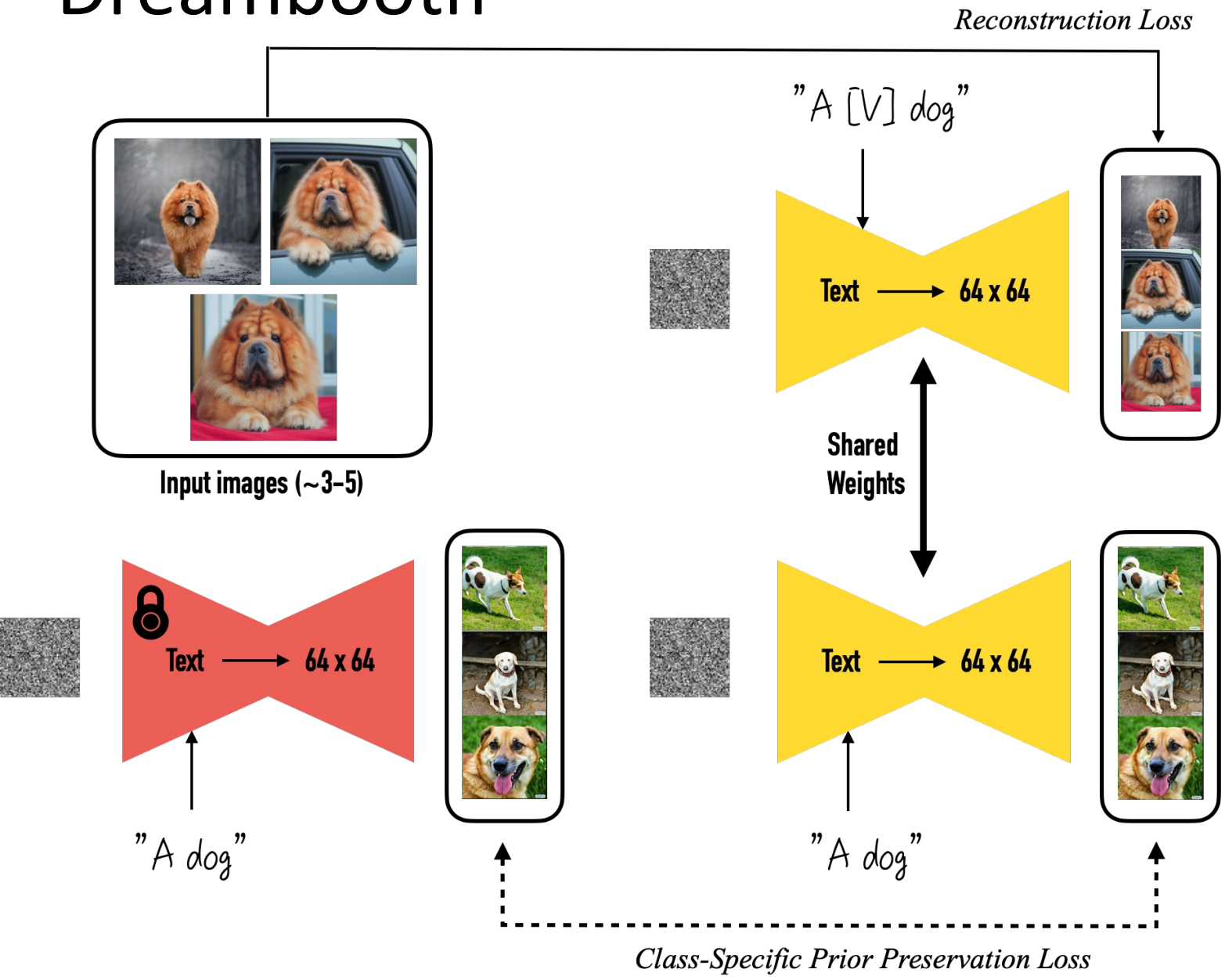


Dreambooth

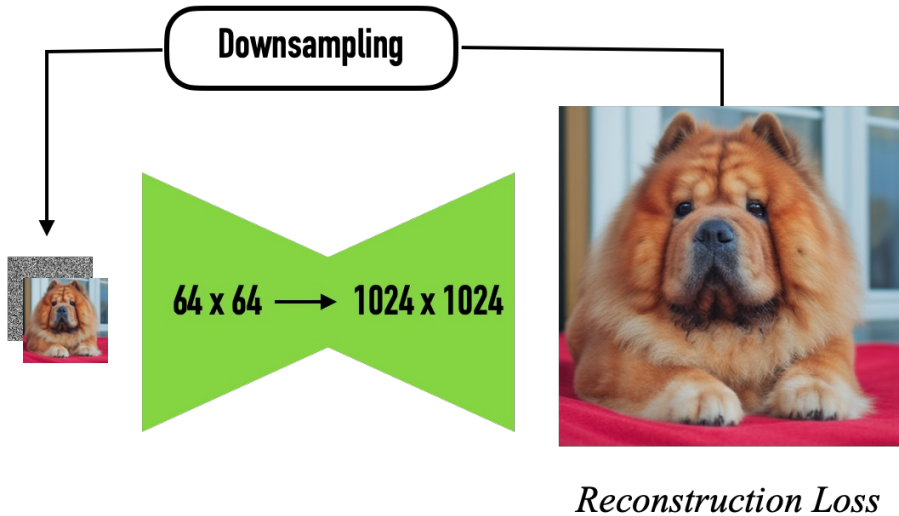

Fine-Tuning
Input
Images (~3-5) +
subject's class name
+ "dog"
Text-to-Image
Pretraind
Text-to-Image
model
DreamBooth
Output
Unique
identifier
"[V]"
Personalized Text-to-Image
Personalized
Text-to-Image
model


Inference
"A [V] dog in
the beach"
Personalized Text-to-Image
"A [V] dog
walking on a
colorful carpet"
Personalized Text-to-Image

Dreambooth



Super-Resolution components:
Fine tuning + unconditional sampling in inference



LoRA : Low-Rank Adaptation

< LoRA의 훈련 과정 >

- Pre-trained weight를 고정된 상태(freeze)로 유지한다.
- Adaptation 중 dense layer의 변화에 대한 rank decomposition matrices를 최적화한다.
- 이를 통해 dense(fc) layer를 간접적으로 훈련 시킨다.

$$h = W_0 x + \Delta W x = W_0 x + BAx$$

$$W_0 \in \mathbb{R}^{d \times k}$$

$$B \in \mathbb{R}^{d \times r}, A \in \mathbb{R}^{r \times k}$$

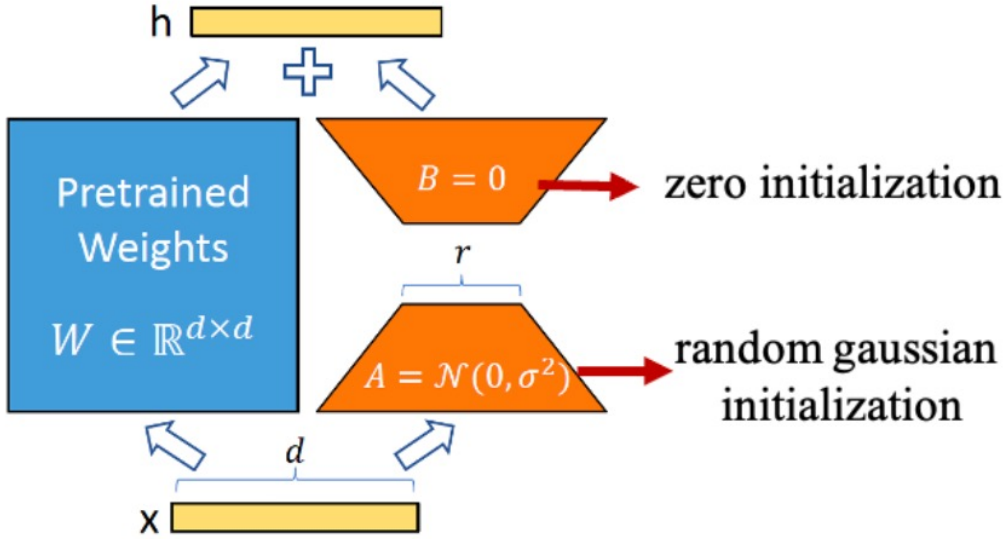
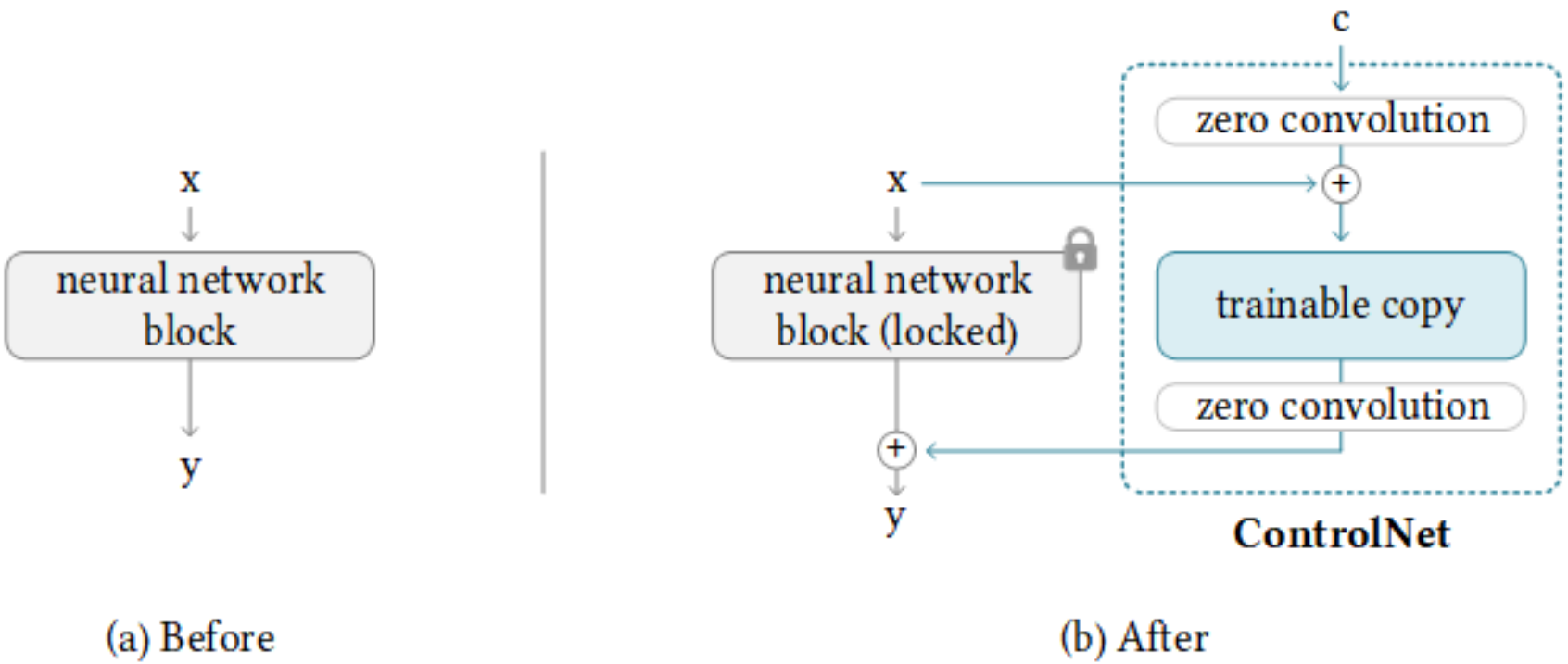
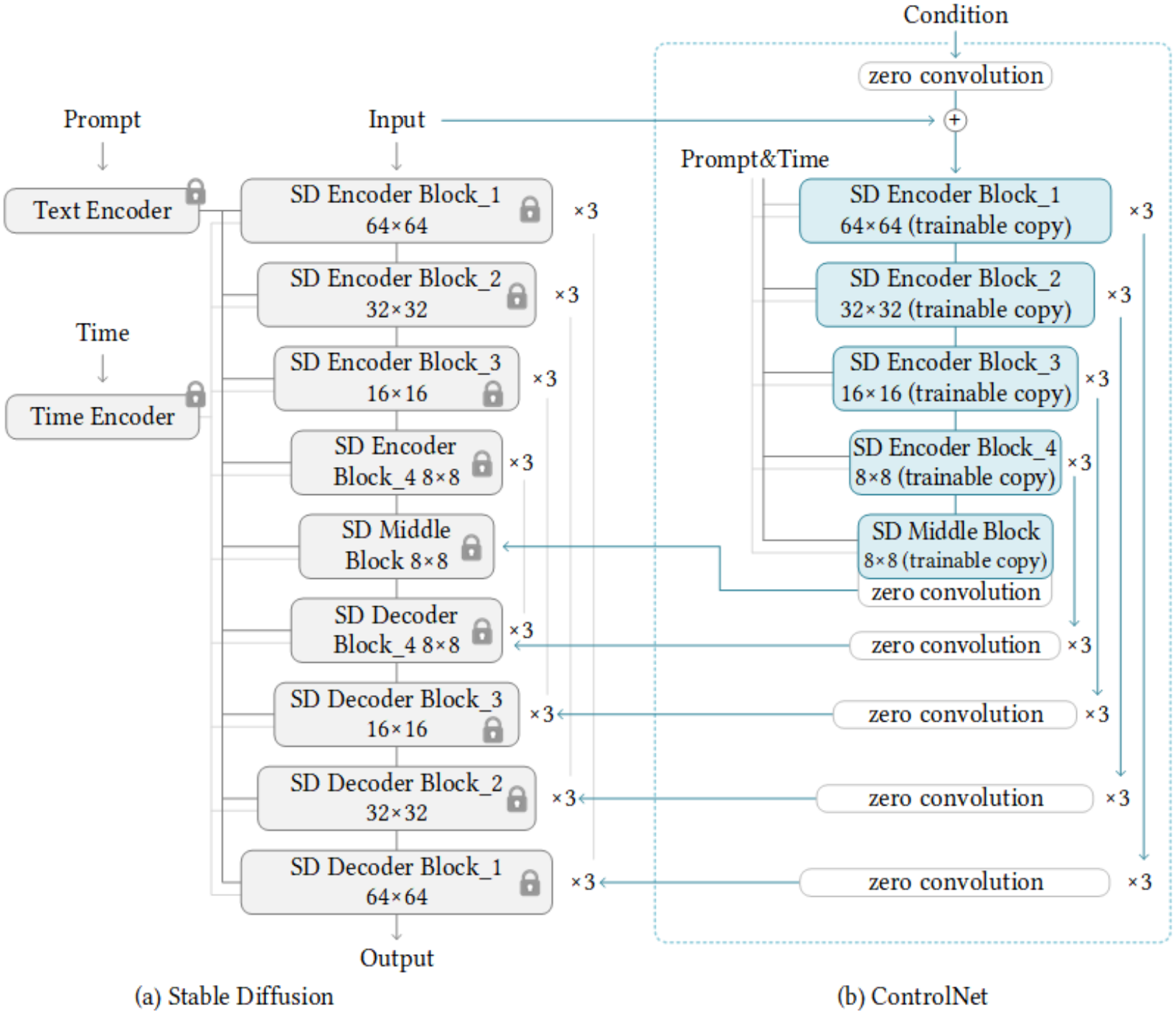


Figure 1: Our reparametrization. We only train A and B.

ControlNet


x
neural network block
y
c
zero convolution
x
neural network block (locked)
trainable copy
zero convolution
y
ControlNet
(a) Before
(b) After

ControlNet


Condition
zero convolution
Prompt
Input
Prompt&Time
Text Encoder
SD Encoder Block_1 64×64
SD Encoder Block_1 64×64 (trainable copy)
×3
SD Encoder Block_2 32×32
SD Encoder Block_2 32×32 (trainable copy)
Time
SD Encoder Block_3 16×16
SD Encoder Block_3 16×16 (trainable copy)
Time Encoder
SD Encoder Block_4 8×8
SD Encoder Block_4 8×8 (trainable copy)
SD Middle Block 8×8
SD Middle Block 8×8 (trainable copy)
zero convolution
SD Decoder Block_4 8×8
SD Decoder Block_3 16×16
SD Decoder Block_2 32×32
SD Decoder Block_1 64×64
Output
(a) Stable Diffusion
(b) ControlNet