값비싼 Diffusion model을 받드는 저비용 MLOps

김태훈
SYMBIOTE AI

이동익
Saige Research

NAVER DEVIEW 2023

김태훈 (carpedm20)

2016

Devsisters

김태훈 (carpedm20)

2016

Devsisters

DEVIEW
2017

책 읽어주는 딥러닝:

배우 유인나가 해리포터를 읽어준다면

김태훈 / carpedm20

DEVSISTERS

김태훈 (carpedm20)

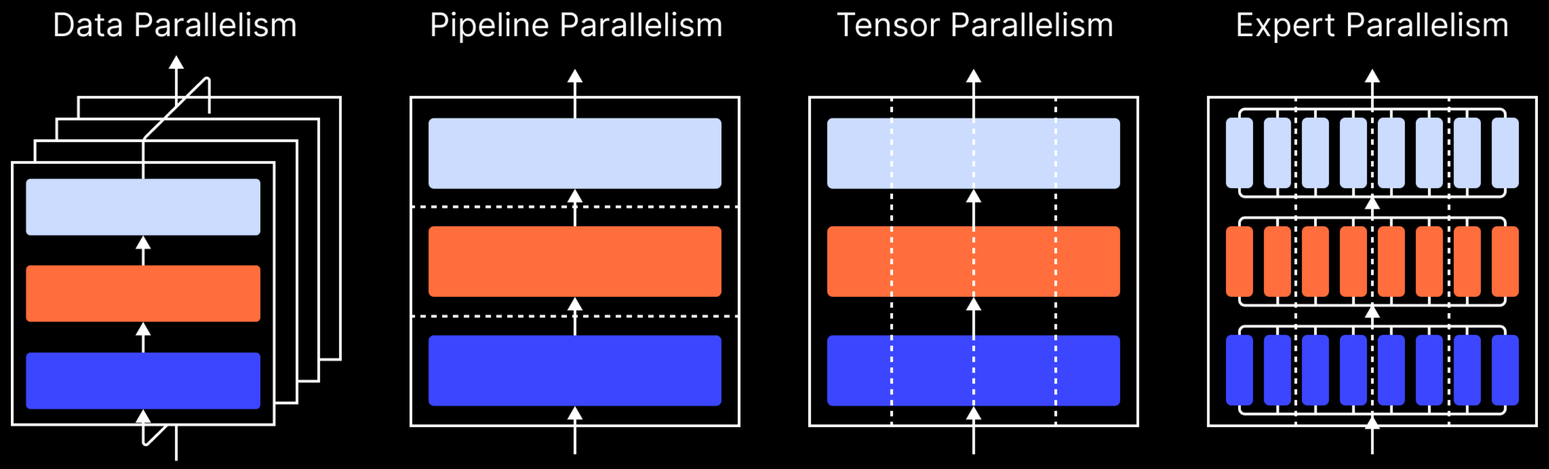
2016

Devsisters

2018

OpenAI

GPT-3, Reinforcement Learning



김태훈 (carpedm20)

2016

Devsisters

2018

2021

SYMBIOTE AI

김태훈 (carpedm20)

2016
Devsisters

2018
OpenAI

2021
SYMBIOTE AI

2D, 3D, and Video generation

값비싼 Diffusion model을 받드는 저비용 MLOps

1. Diffusion model은 무엇인가?

값비싼 Diffusion model을
받드는 저비용 MLOps

값비싼 Diffusion model을

저비용

받드는 MLOps

2. 어떻게 최적화 했는가?

값비싼 Diffusion model을

받드는 저비용 MLOps

3. 어떻게 serving 했는가?

1. Diffusion model은 무엇인가?

Diffusion model

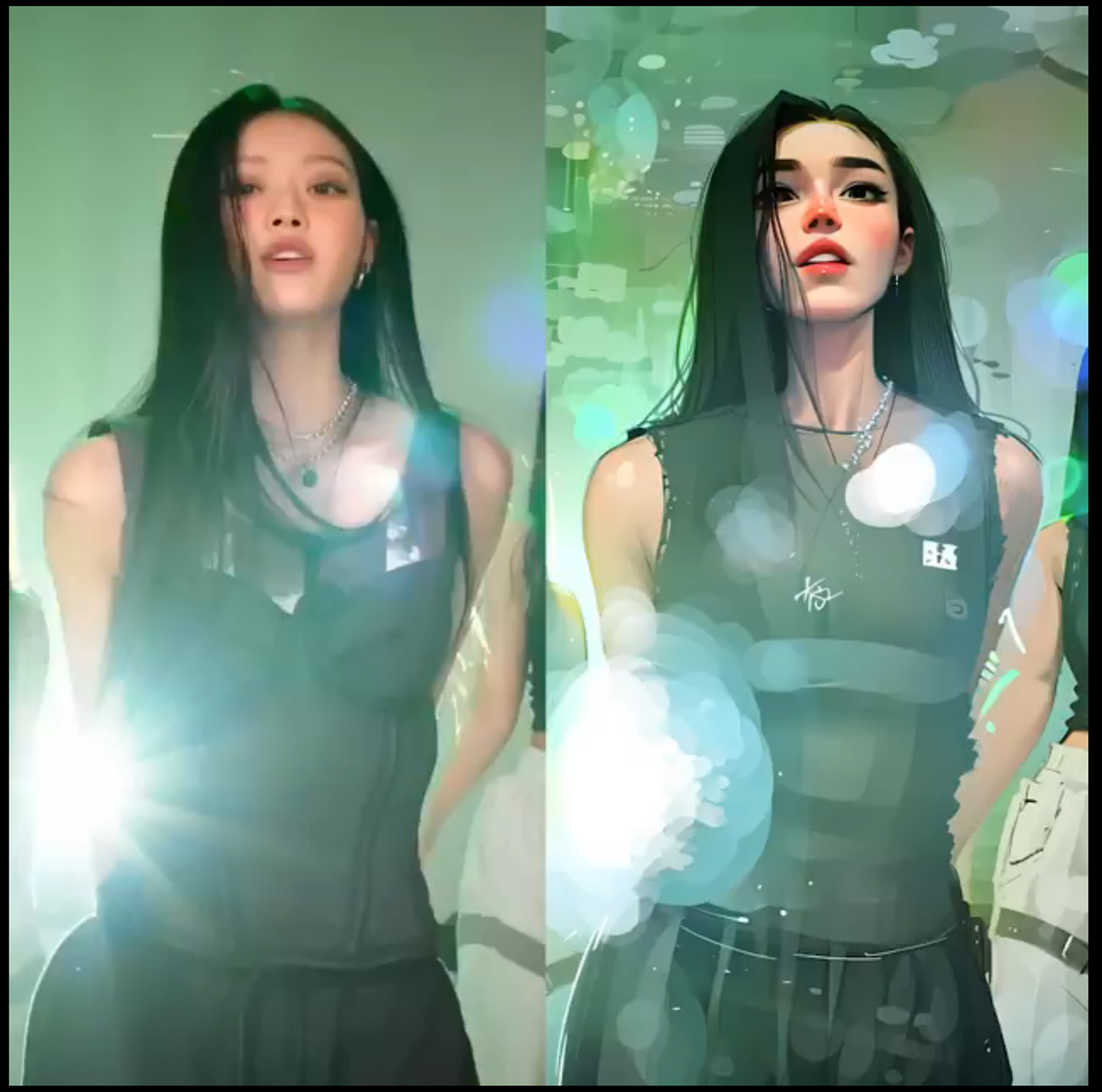
값비싼 Diffusion model를 밟드는 저비용 MLOps

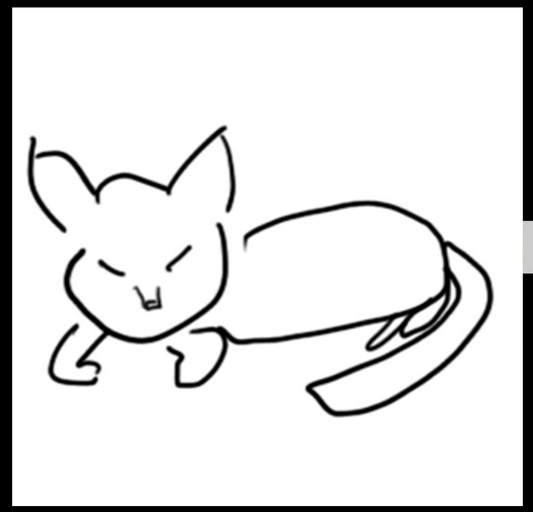
Diffusion model이 할 수 있는 것

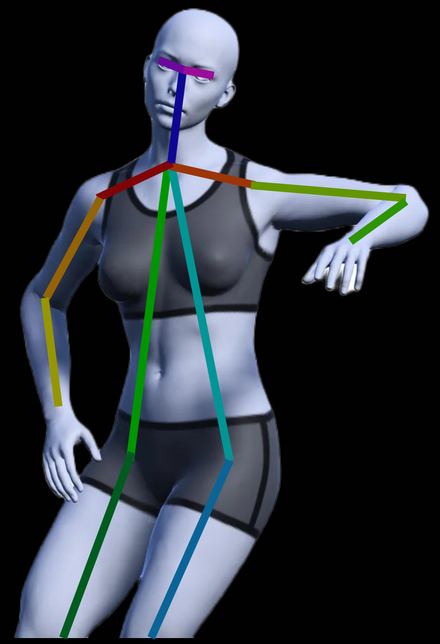
Jeova Sanctus Unus
@Adonai_Jeova

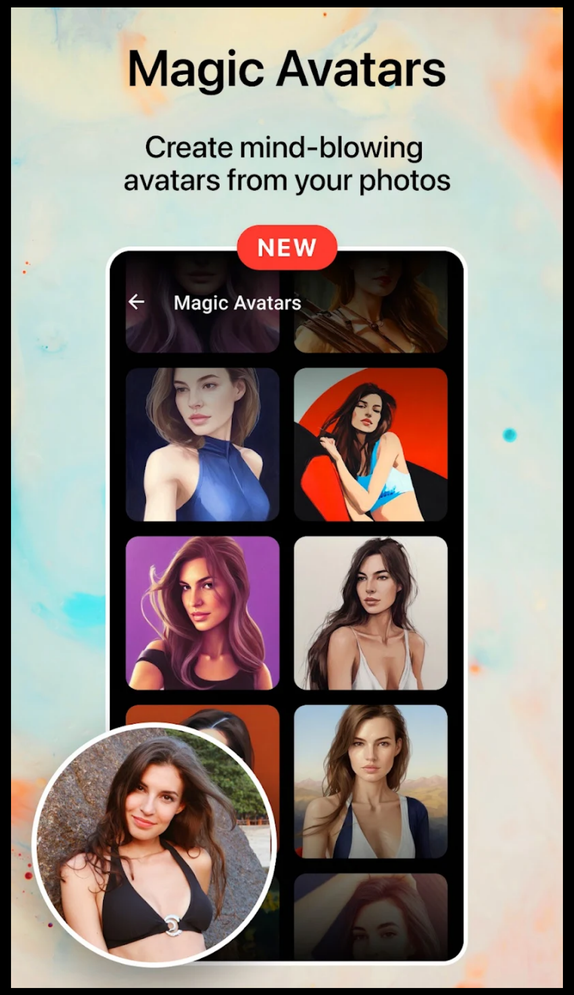
지금 탐라에서 유행하는거.. 뉴턴 돌려봤었음..

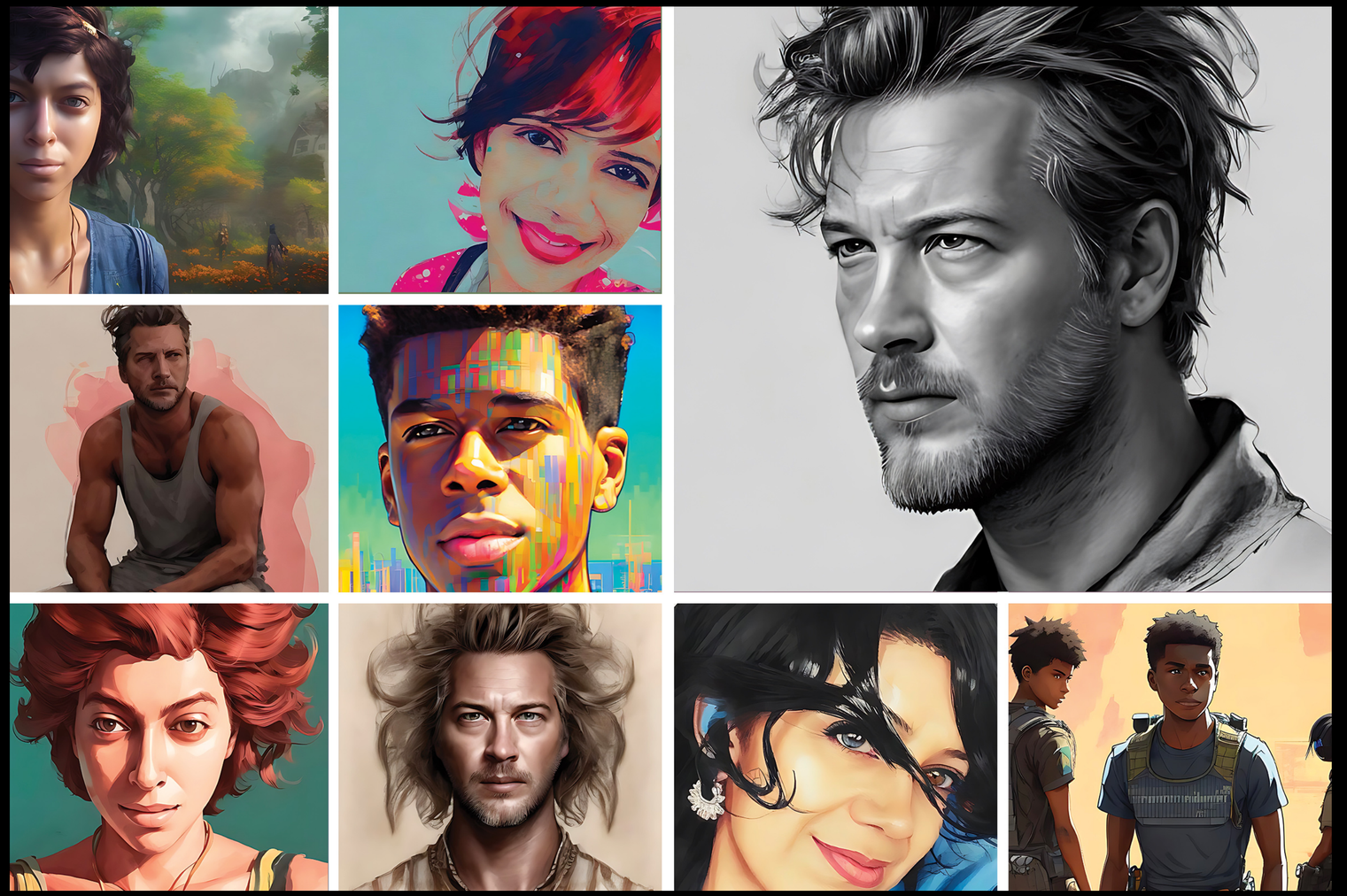
2,947 Retweets **1,325** Quote Tweets **1,915** Likes



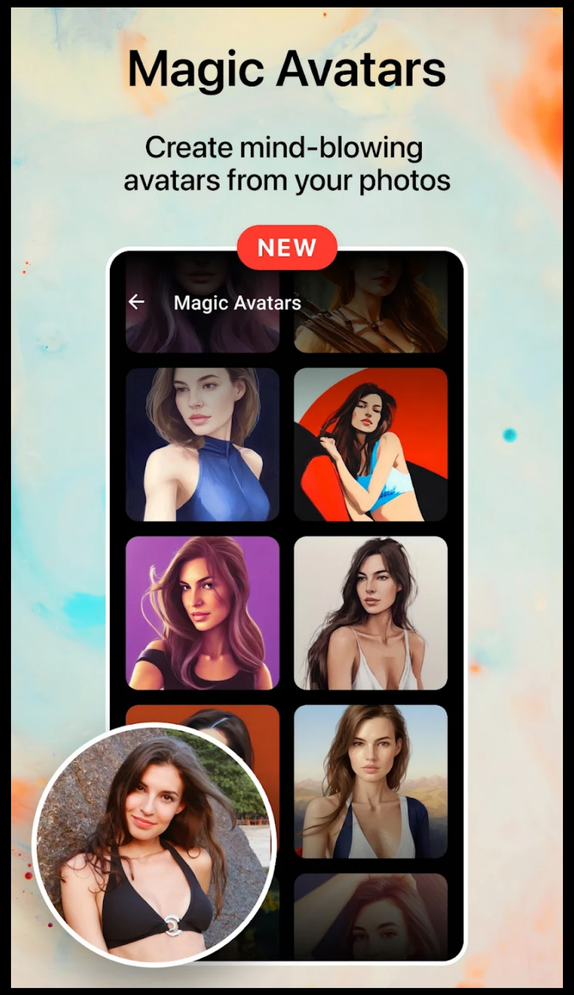








인물 사진을 디지털 아바타로 바꿔주는 앱



출시 후 한 달 매출 400억

1. Diffusion은 무엇이 다른가?

Diffusion model 의 정의

Diffusion Model is

Multimodal Image Generation model with diffusion process

Diffusion Model is

Multimodal

Image Generation model

with diffusion process

Image Generation model

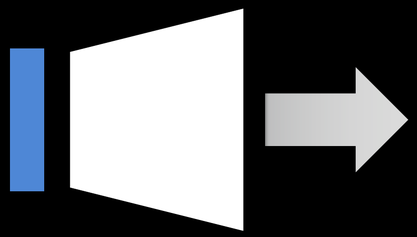
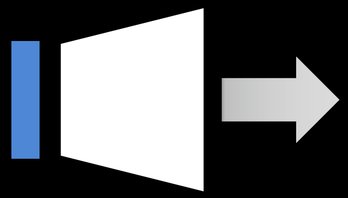


Image Generation model의 역사



BigGAN

2018

StyleGAN2

2020

DALL·E 2

2022

Diffusion Model is

Multimodal

Image Generation model

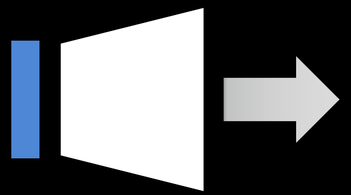
with diffusion process

Diffusion Model is

Multimodal Image Generation model with diffusion process

Singlemodal

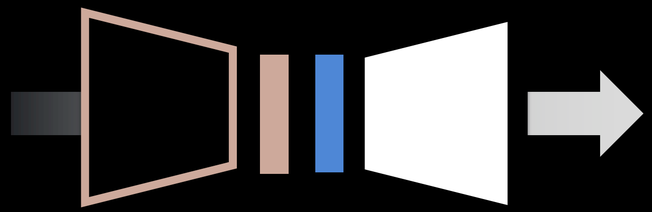
Multimodal



Singlemodal

Text

“말타는
우주 비행사”



Multimodal

GAN

2020

Singlemodal

“우주 수프를 담은 그릇”

“말타는 우주 비행사”

“미친 과학자처럼
실험하고 있는 테디 베어”

Multimodal

Diffusion Model is

Multimodal Image Generation model with diffusion process

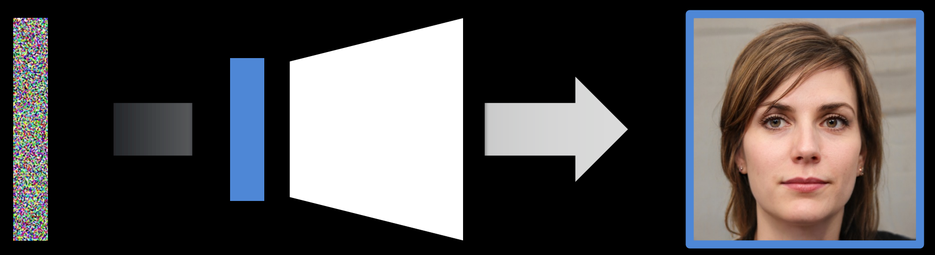
Diffusion Model is

Multimodal

Image Generation model

with diffusion process

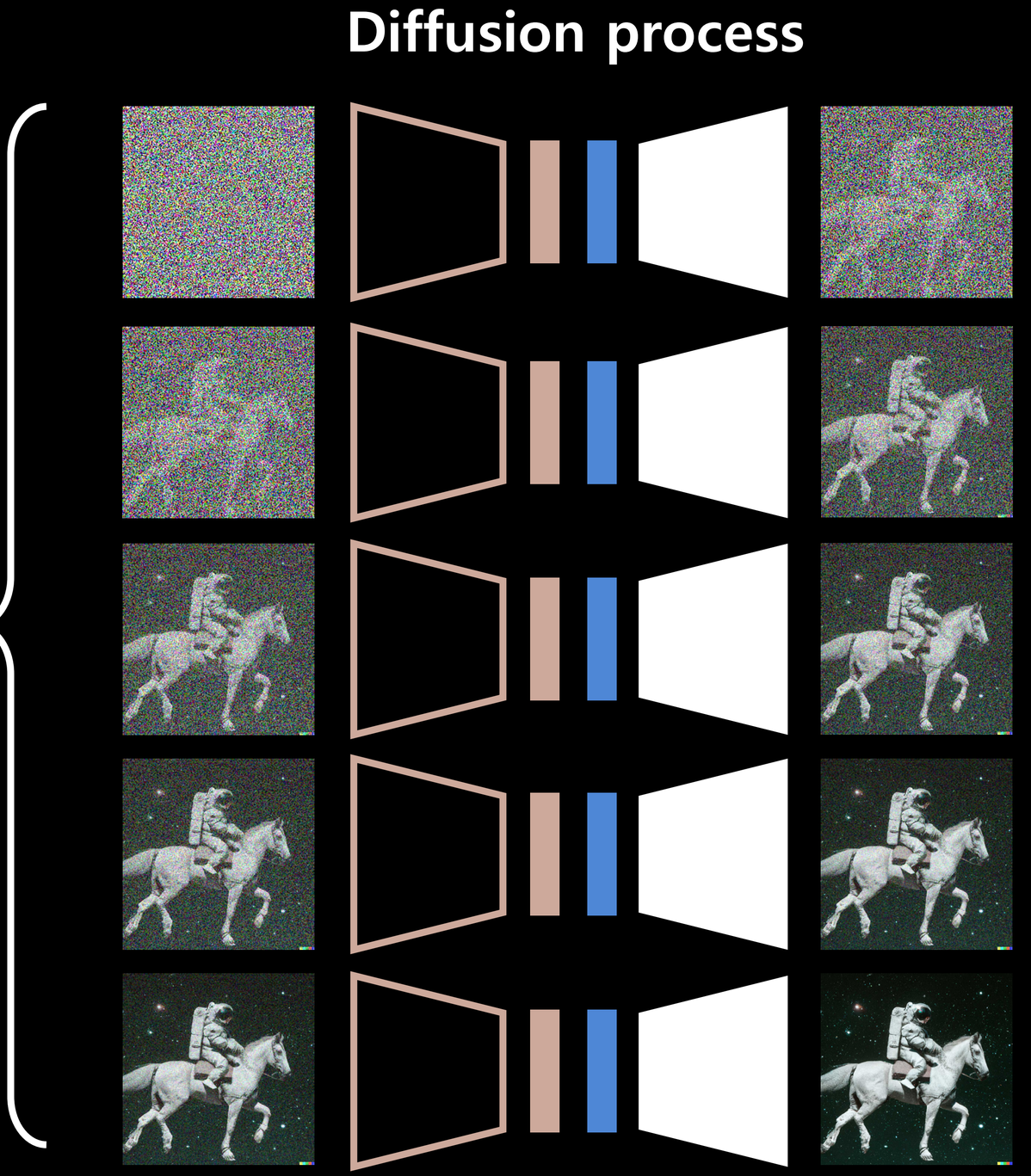
GAN



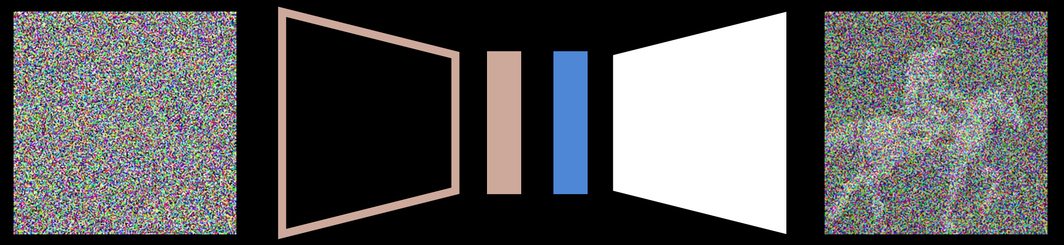
Diffusion process

학습시 1,000번 정도 반복

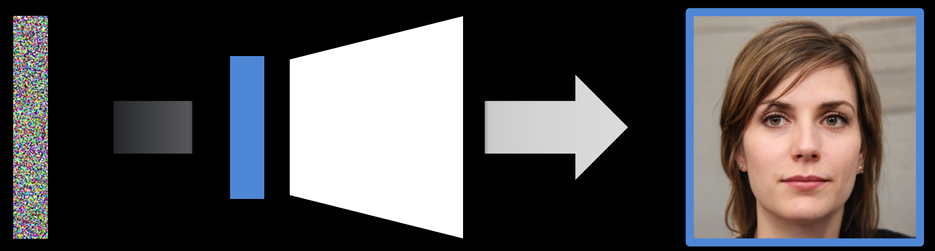
생성할 땐 25 ~ 50 step



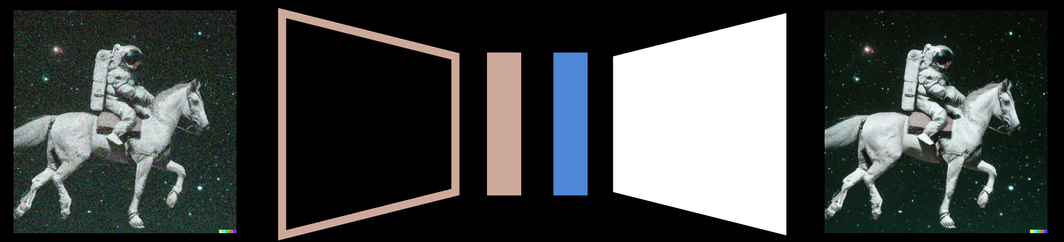
Diffusion process



GAN



× 50번 반복하기
때문에 느리다



Diffusion Model

단점

1. 생성이 느리다

1. 생성이 느리다

= GPU 비용이 많이 든다

2. 학습에 리소스가 많이 든다

2. 학습이 느리다

= GPU 비용이 많이 든다

× 5,700배

2.5%

70,000개의 이미지

2018

1,800개의 이미지

2020

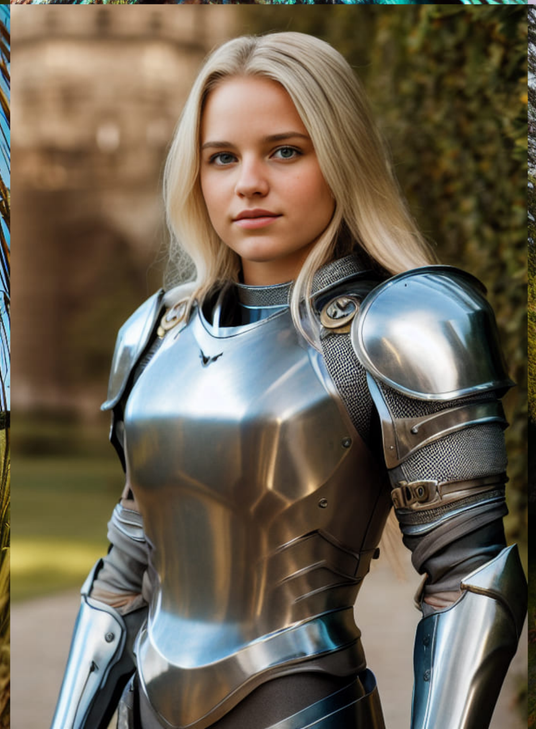
400,000,000개의 이미지

2022







그래서

Diffusion Model is

Multimodal Image Generation model with diffusion process

값비싼 Diffusion model를

밟드는 저비용 MLOps

값비싼 Diffusion model를

저비용

밭드는 MLOps

1. Diffusion은 무엇이 다른가?

Diffusion model은
그동안 어떻게 발전해 왔는가?

Stable Diffusion

DreamBooth & LoRA

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

지난 6개월 간의 발전

2021.12 ~ 2022.08

Stable Diffusion

DreamBooth & LoRA

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

몇몇 연구자들 사이에서

Diffusion model이 엄청 좋아졌고

2021.12 ~ 2022.08

Stable Diffusion

DreamBooth & LoRA

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

몇몇 연구자들 사이에서

Diffusion model이 엄청 좋아졌고

몇 억 정도면 만들 수 있겠다는

생각을 하게 되었다

2021.12 ~ 2022.08

Stable Diffusion

DreamBooth & LoRA

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

stability.ai

runway

EleutherAI

LAION

2021.12 ~ 2022.08

Stable Diffusion

DreamBooth & LoRA

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

150,000 A100 GPU Hours

= $600,000

2021.12 ~ 2022.08

Stable Diffusion

DreamBooth & LoRA

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

stability.ai

API News FAQ English

Stable Diffusion Public Release

It is our pleasure to announce the public release of stable diffusion following our release for researchers [https://stability.ai/blog/stable-diffusion-announcement]

2021.12 ~ 2022.08

Stable Diffusion

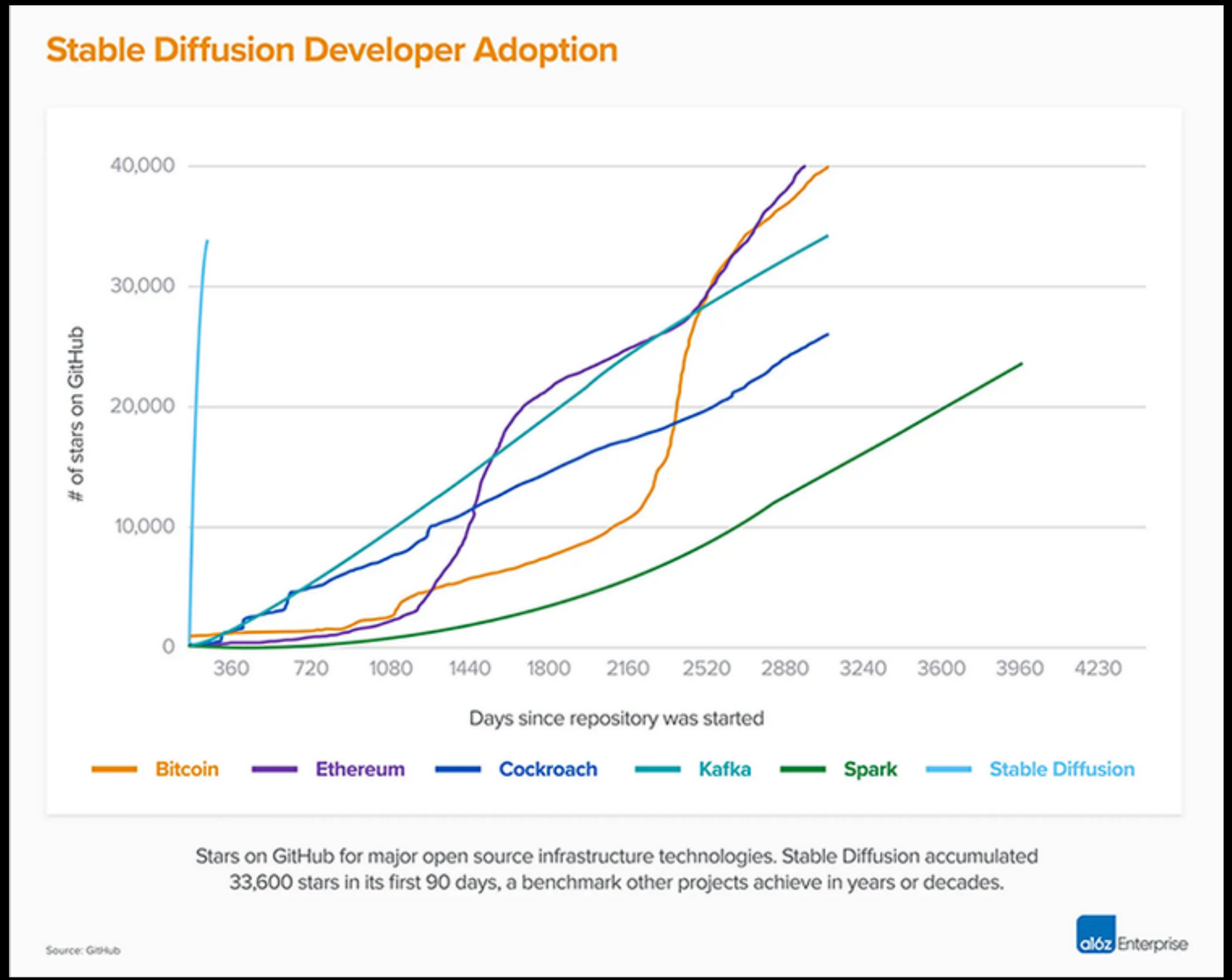
DreamBooth & LoRA

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet



2021.12 ~ 2022.08

Stable Diffusion

DreamBooth & LoRA

xformers

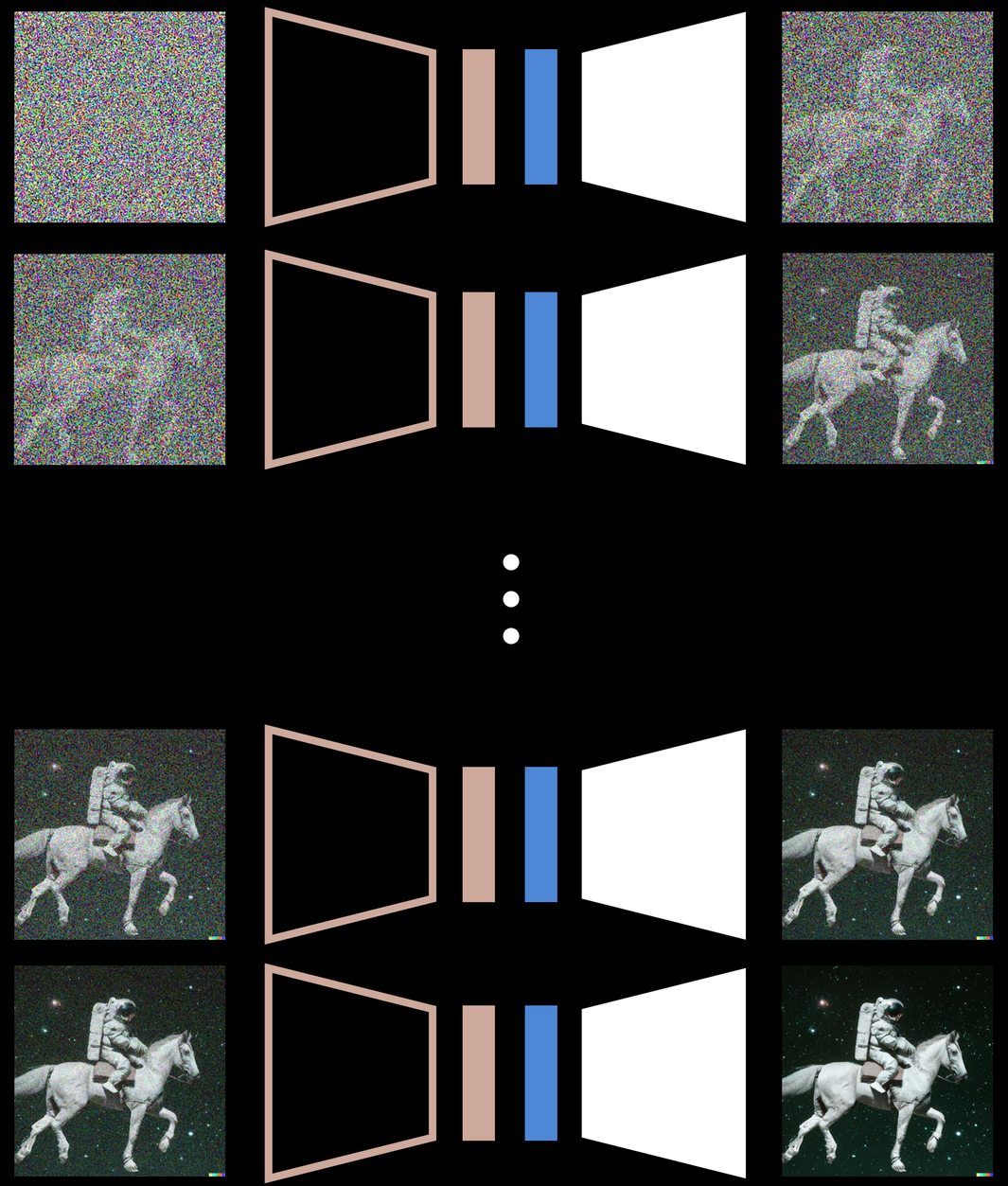
Prompt-to-Prompt

InstructPix2Pix

ControlNet

50 step

V100 기준 10초



2021.12 ~ 2022.08

Stable Diffusion

DreamBooth & LoRA

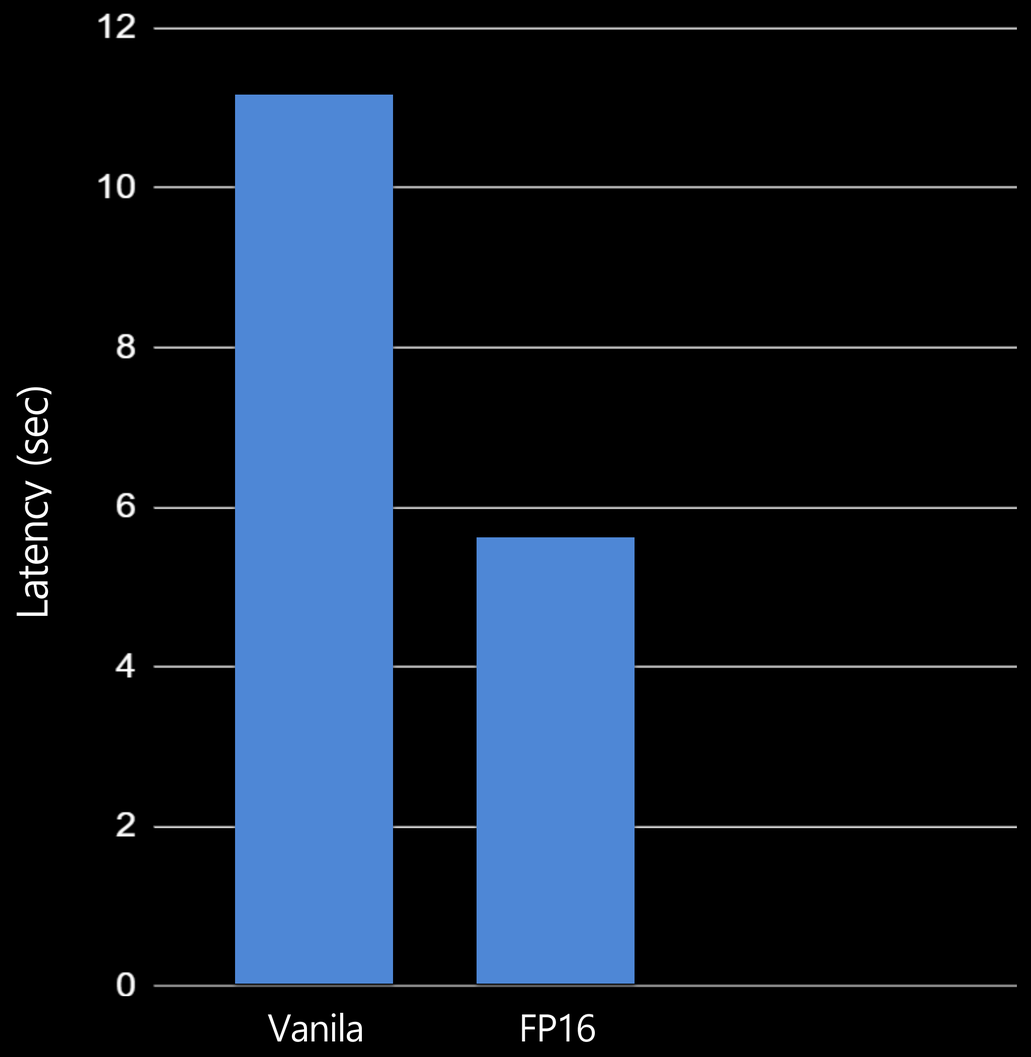
xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

Latency



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation

Nataniel Ruiz Yuanzhen Li Varun Jampani Yael Pritch Michael Rubinstein Kfir Aberman

Google Research

2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

4억 개의 이미지 없이도

새로운 모델을 학습할 수 있을까?

2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

5 ~ 20개의 학습 이미지

“강아지X”

“로마에 간 강아지X”

“머리 깎는 강아지X”

2021.12 ~ 2022.08

Stable Diffusion

2022.08

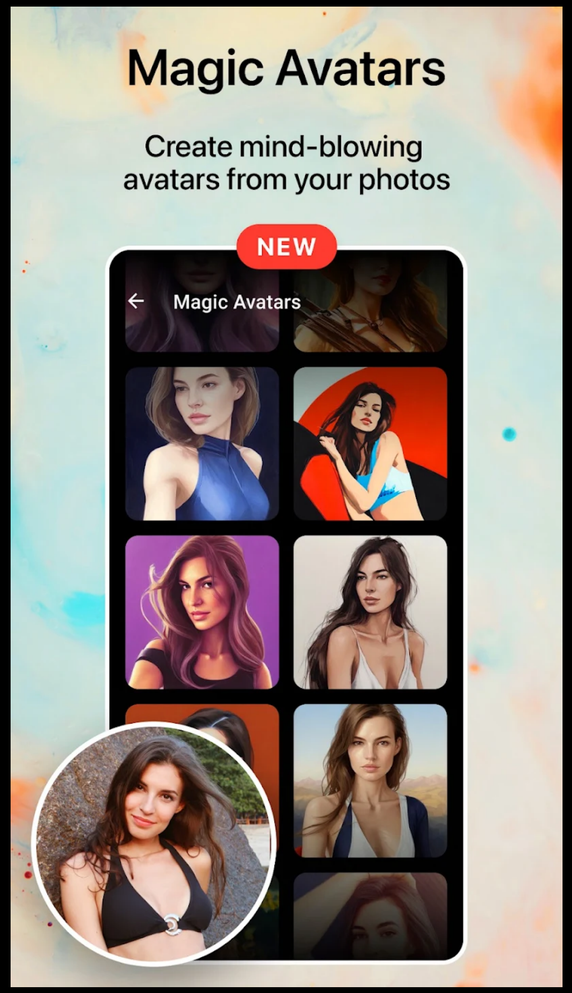
DreamBooth & LoRA

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet



2022.11

LENSA

개발 기간 3개월

2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

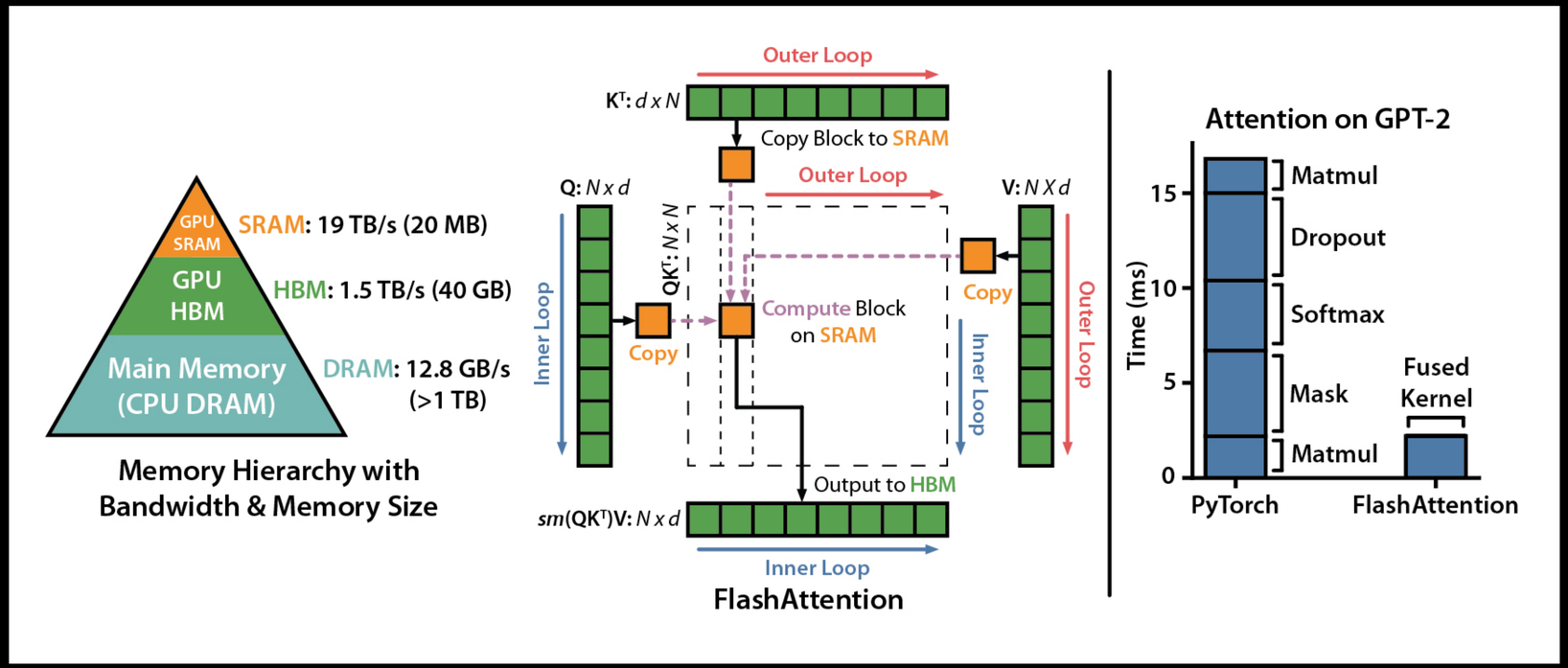
Prompt-to-Prompt

InstructPix2Pix

ControlNet

FlashAttention

Fast and Memory-Efficient Exact Attention with IO-Awareness



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

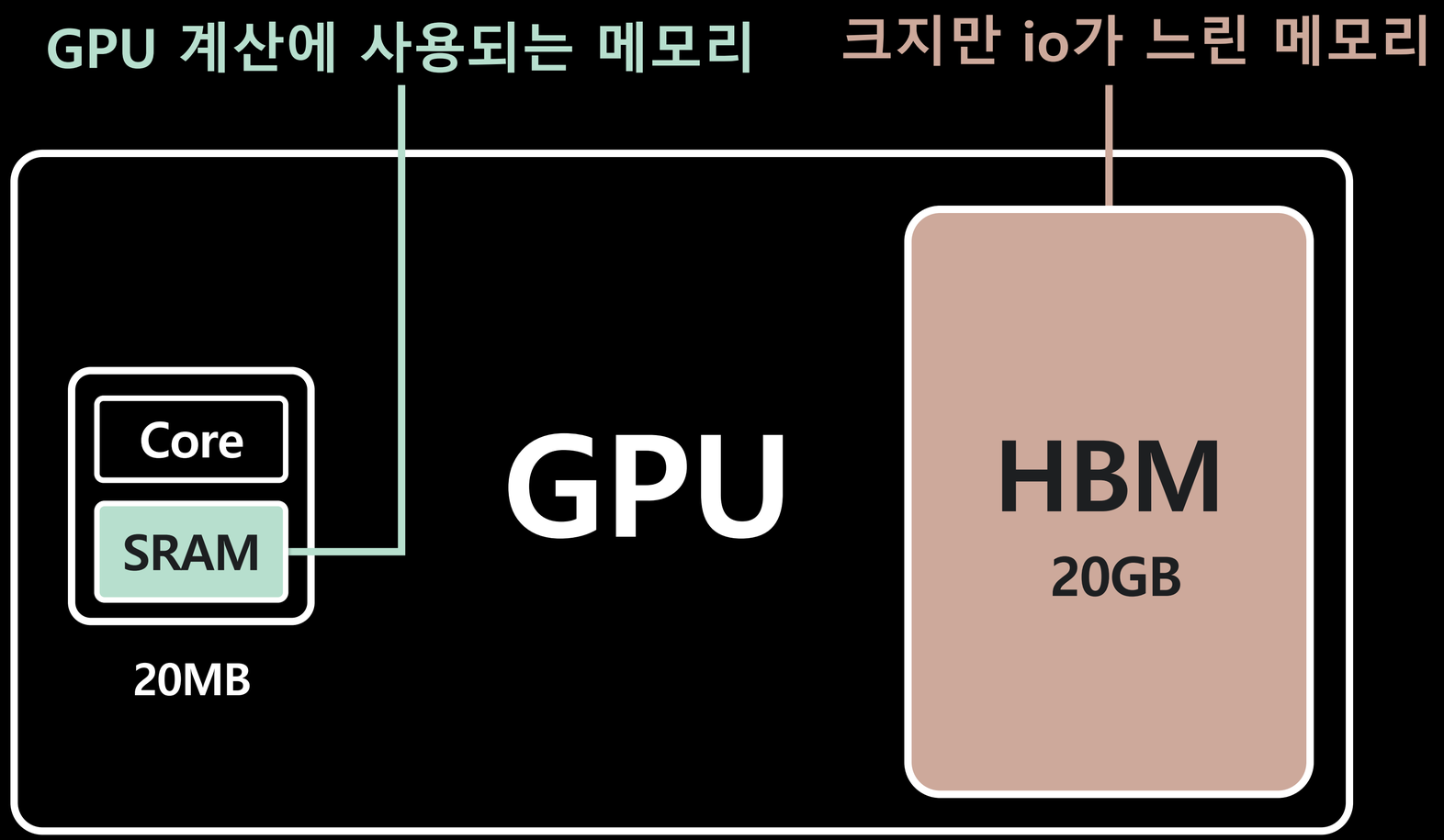
xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

FlashAttention



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

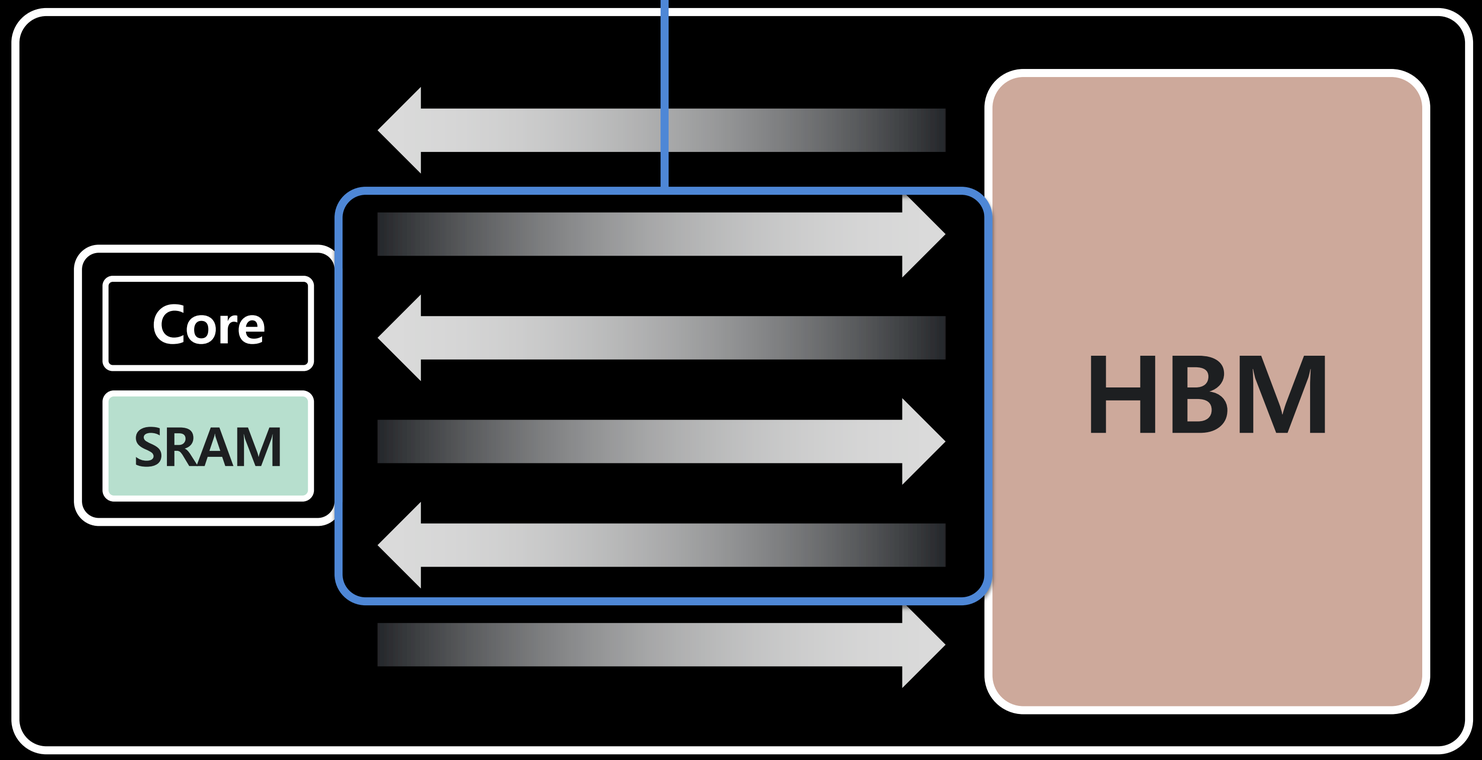
Prompt-to-Prompt

InstructPix2Pix

ControlNet

FlashAttention

다수의 I/O 과정을 없애는 방법



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

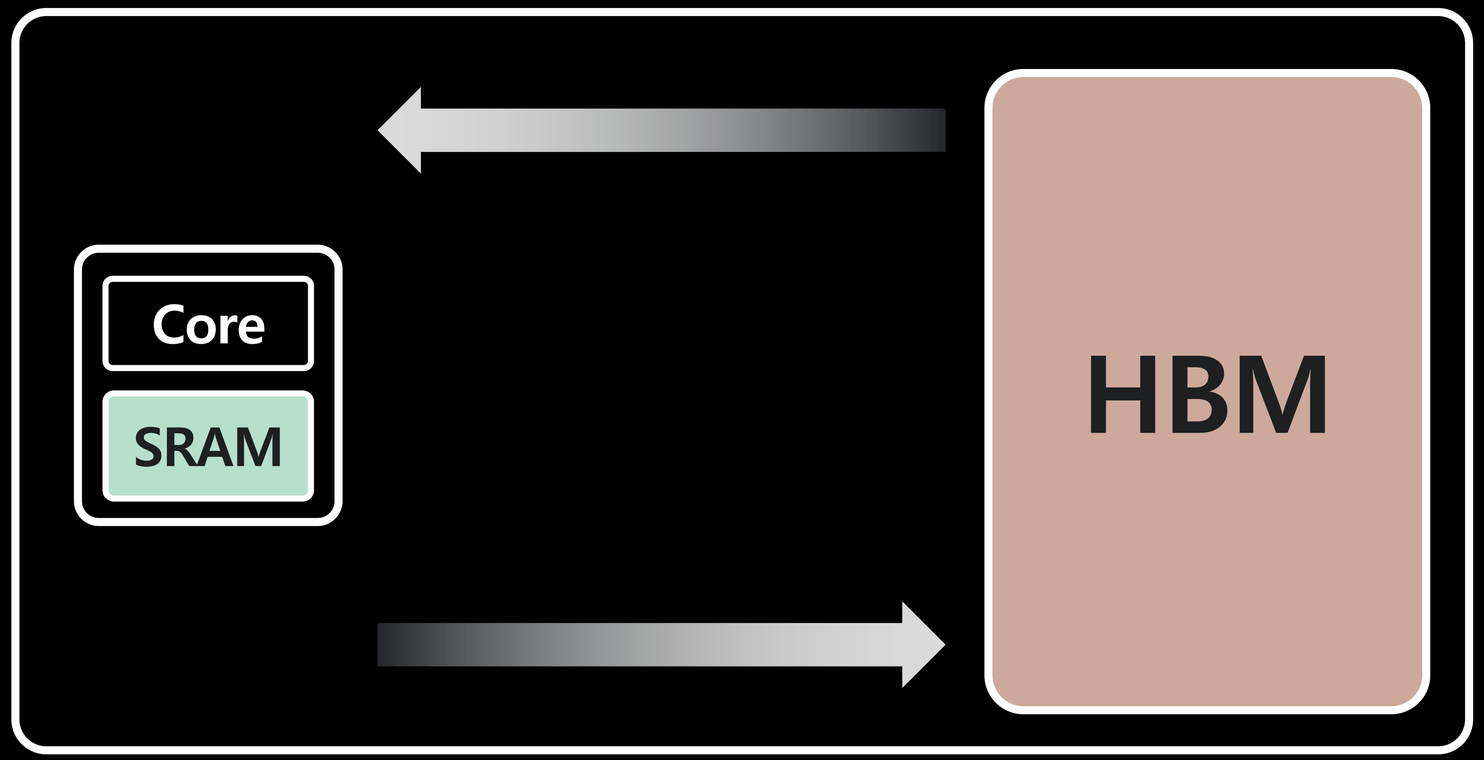
xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

FlashAttention



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

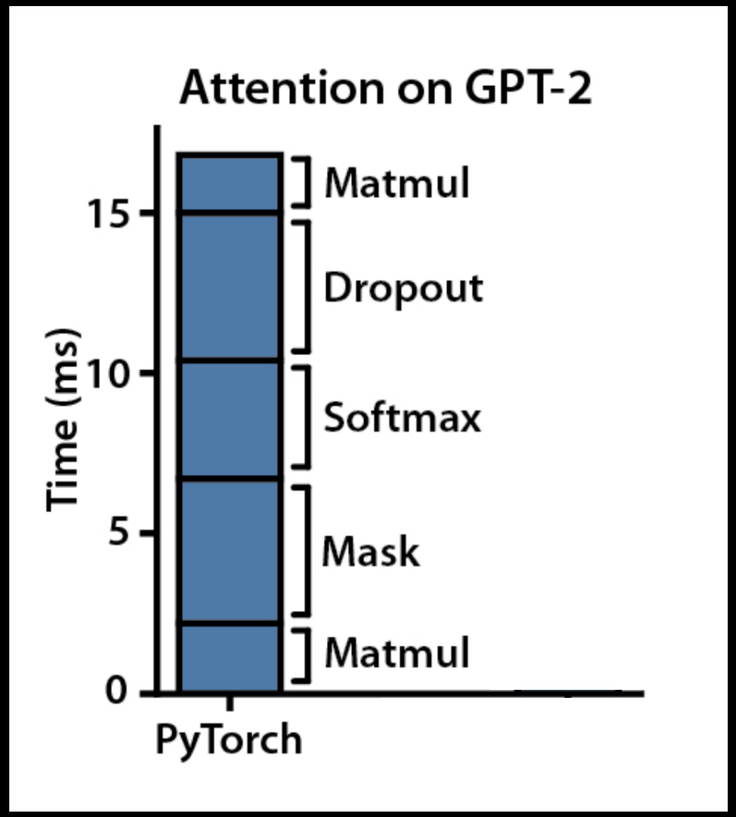
2022.09

xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

Up to 2x speedup on GPUs using memory efficient attention #532

Code

Merged

patil-suraj merged 15 commits into huggingface:main from MatthieuTPHR:memory_efficient_attention on Nov 2, 2022

Prompt-to-Prompt

InstructPix2Pix

ControlNet

2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

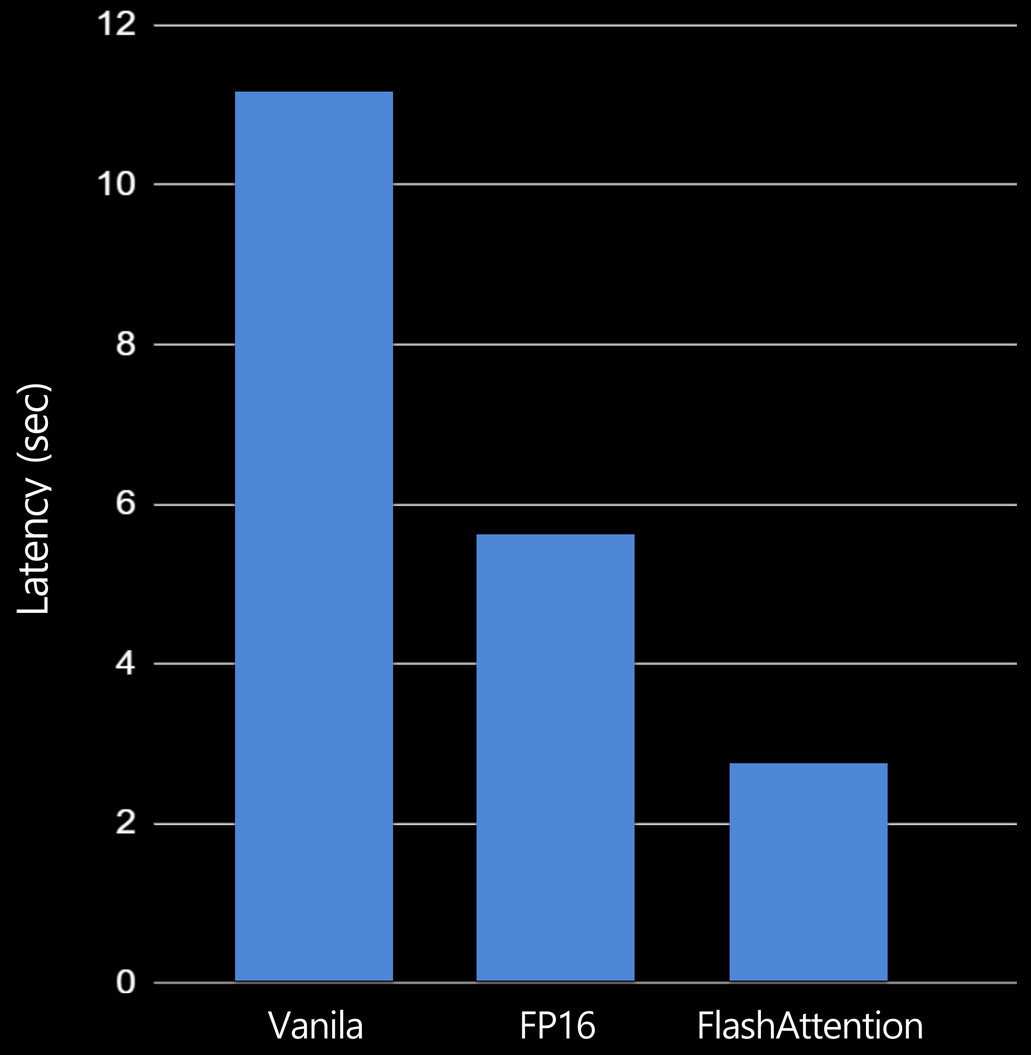
xformers

Prompt-to-Prompt

InstructPix2Pix

ControlNet

Latency



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

2022.10

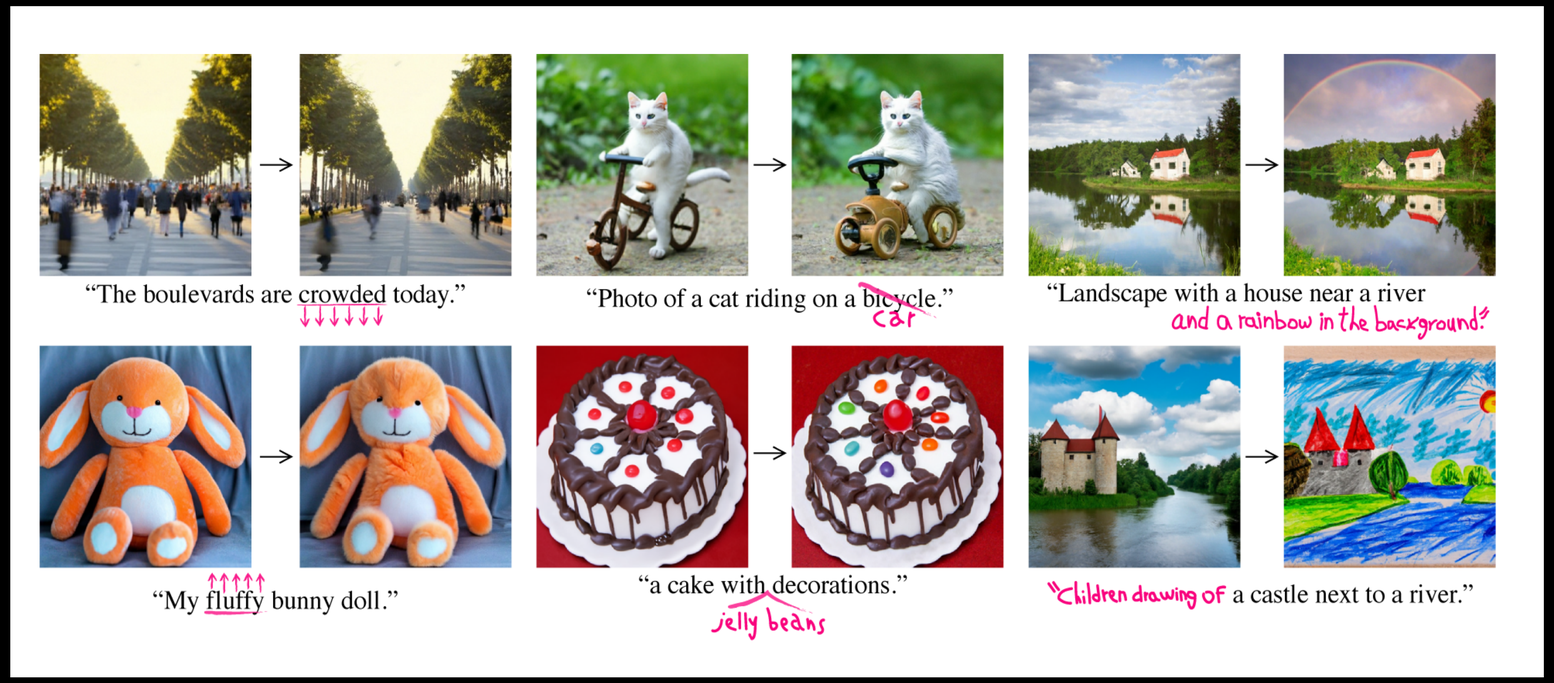
Prompt-to-Prompt

InstructPix2Pix

ControlNet

Prompt-to-Prompt Image Editing with Cross-Attention Control

[1] Google Research [2] Tel Aviv University



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

2022.10

Prompt-to-Prompt

InstructPix2Pix

ControlNet

Prompt-to-Prompt

Image Editing with Cross-Attention Control

"소파에 누워있는 아이"

"소파에 누워있는 **노란 옷을 입은 아이**"

"... **금발의** 아이"

"... **눈을 감는** 아이"

2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

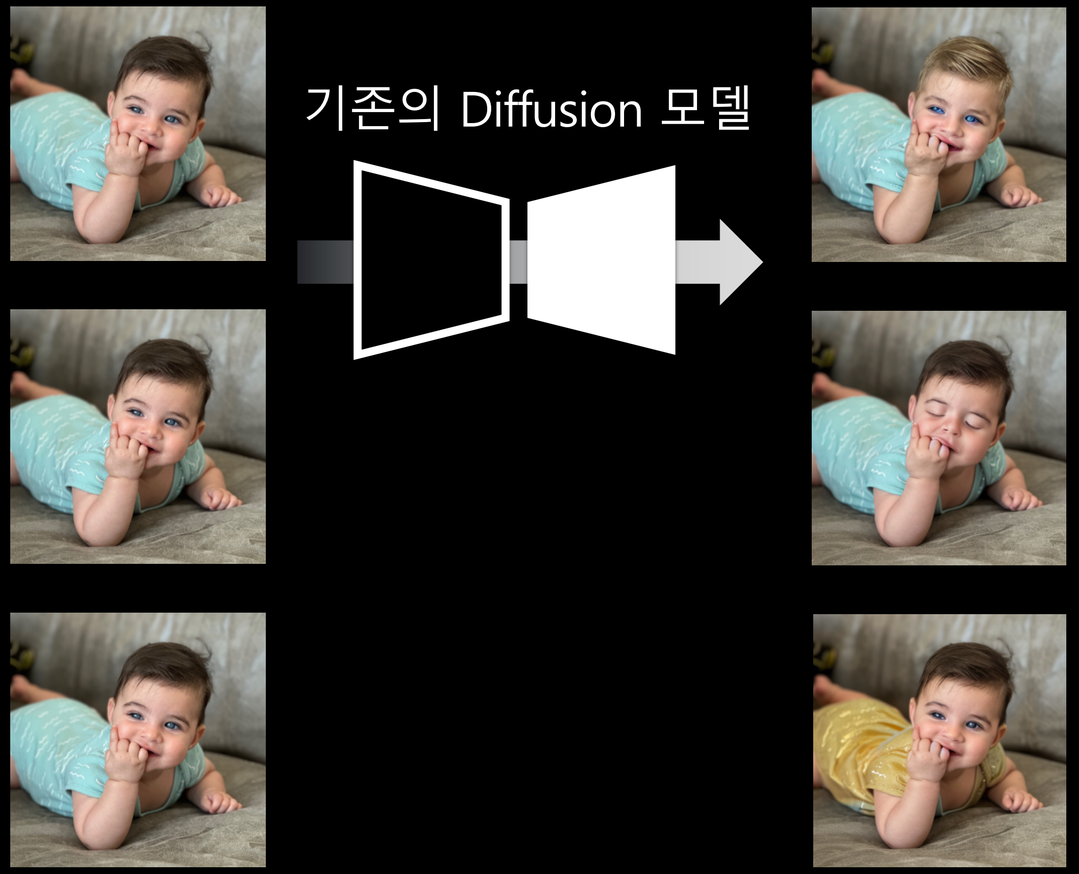
2022.10

Prompt-to-Prompt

InstructPix2Pix

ControlNet

학습 데이터를 만들 때도 사용 가능



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

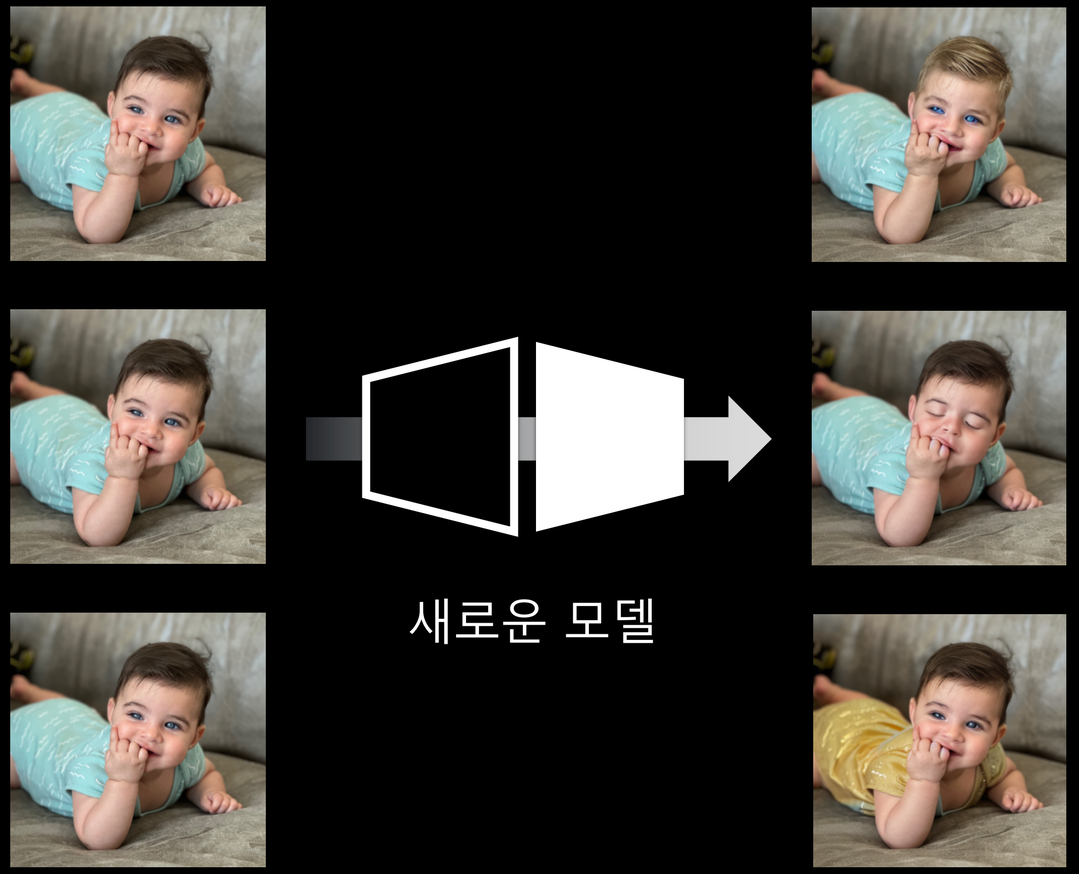
2022.10

Prompt-to-Prompt

InstructPix2Pix

ControlNet

학습 데이터를 만들 때도 사용 가능



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

2022.10

Prompt-to-Prompt

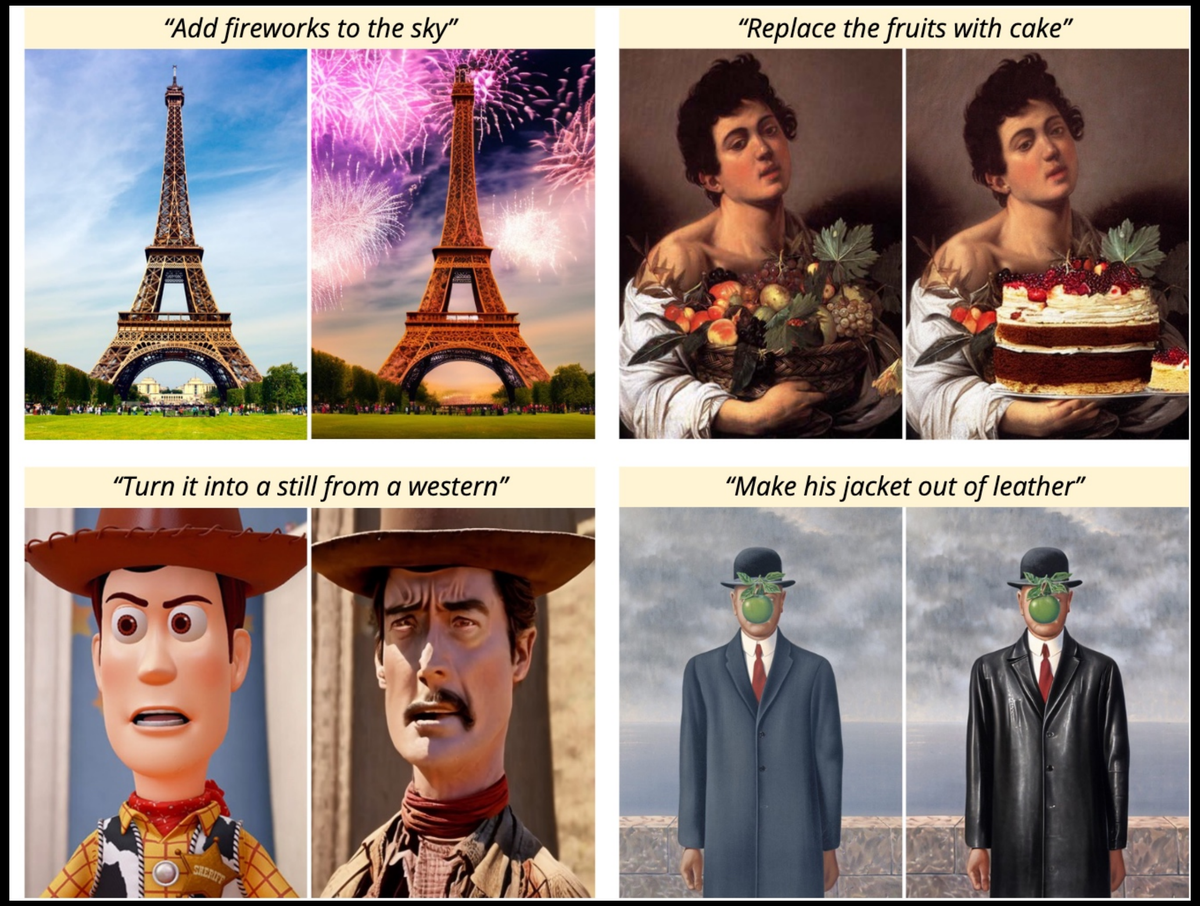
2022.12

InstructPix2Pix

ControlNet

InstructPix2Pix

Learning to Follow Image Editing Instructions



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

2022.10

Prompt-to-Prompt

2022.12

InstructPix2Pix

ControlNet

"소파에 누워있는 아이"

"소파에 누워있는 **금발** 아이"

GPT와 대화하 듯 **"금발로 만들어 줘"**

라는 명령어로 수정하려면?

2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

2022.10

Prompt-to-Prompt

2022.12

InstructPix2Pix

ControlNet

"쿠키가 든 바구니"

"고양이"

"소파에 누워있는 아이"

"사과가 든 바구니"

"썬글라쓰를 쓰고 있는 고양이"

"소파에 누워있는 금발 아이"

GPT3

"금발로 만들어 줘"

"썬글라쓰를 씌워 줘"

"사과로 바꿔줘"

2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

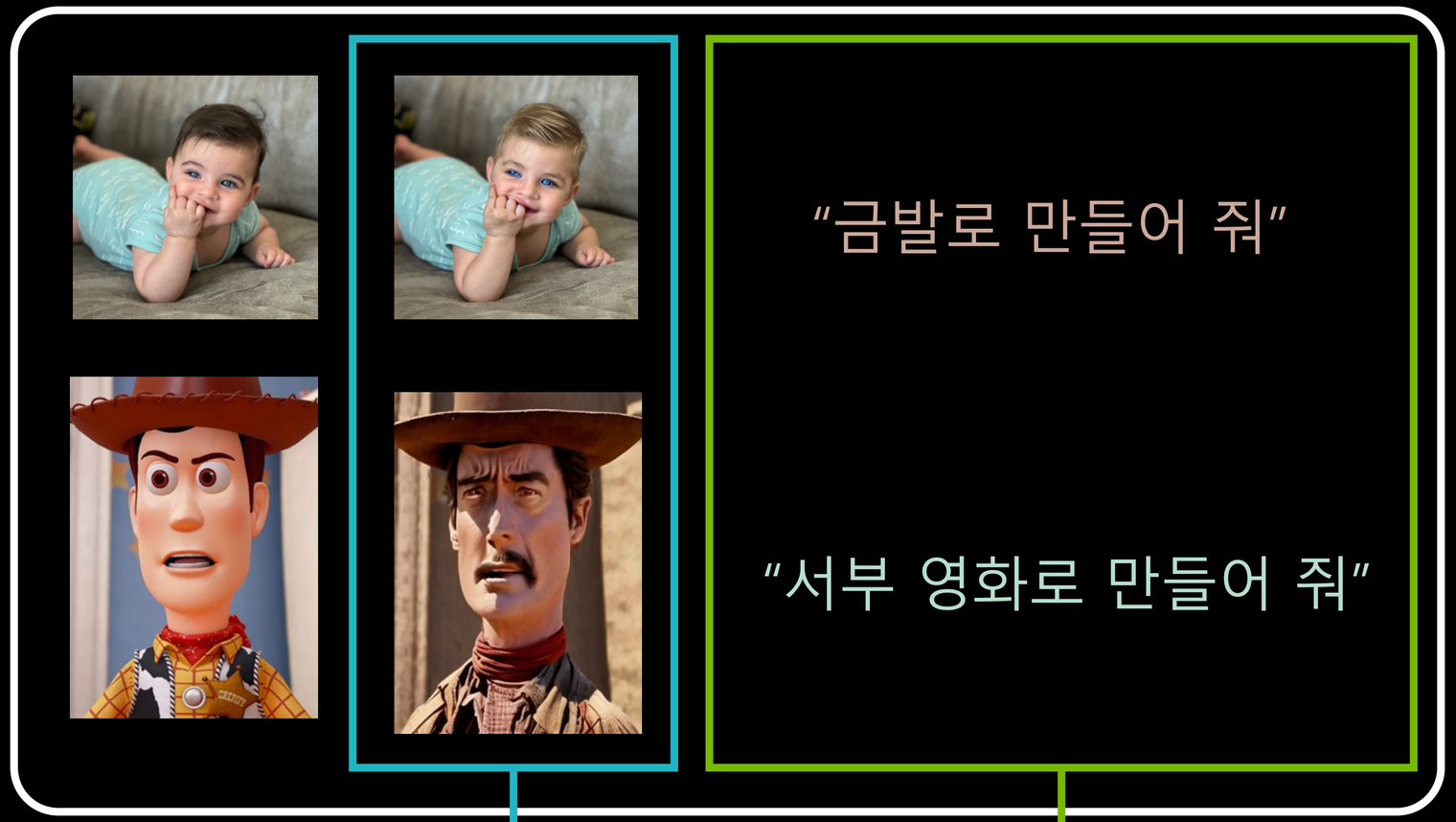
2022.10

Prompt-to-Prompt

2022.12

InstructPix2Pix

ControlNet



2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

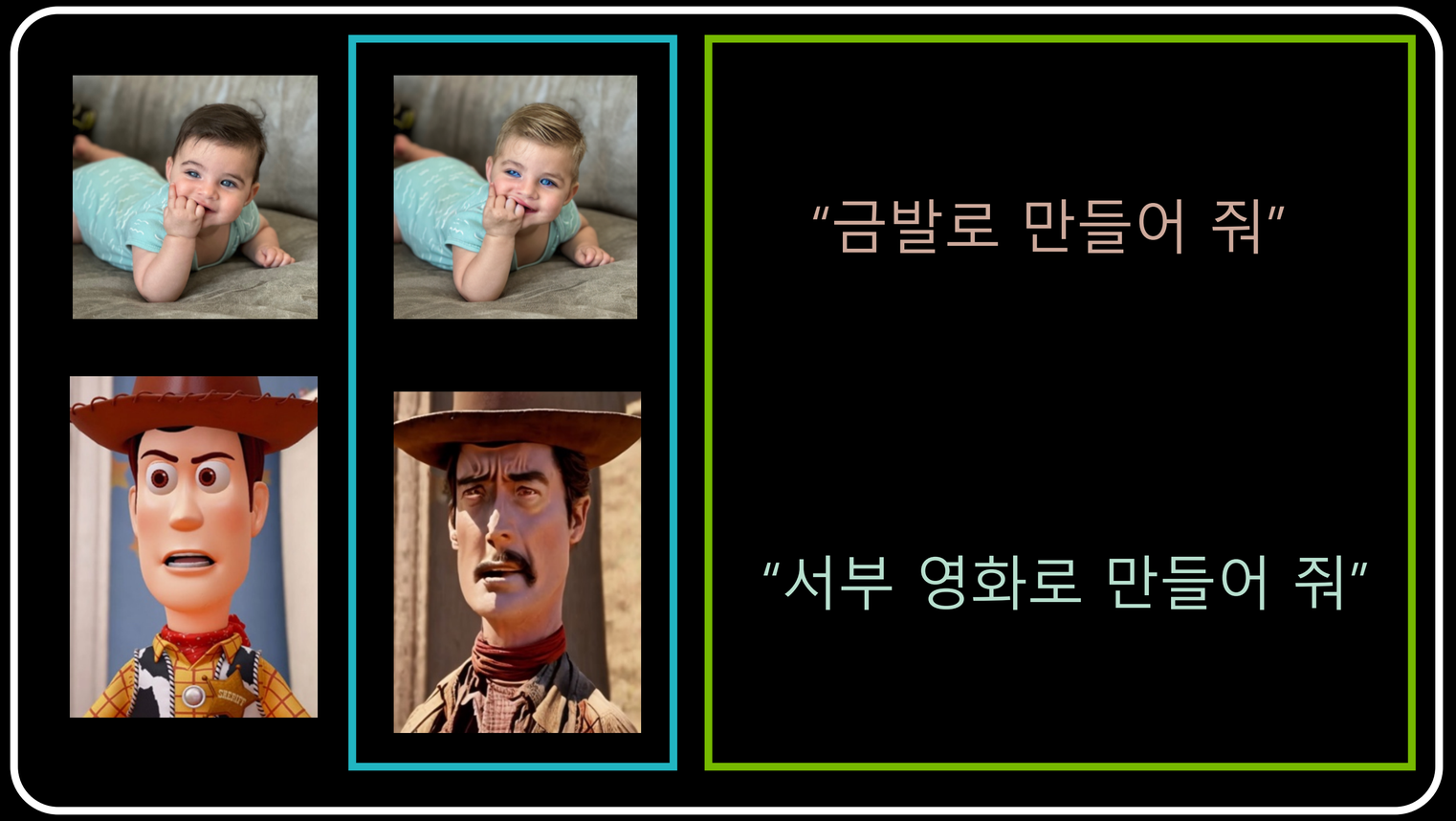
2022.10

Prompt-to-Prompt

2022.12

InstructPix2Pix

ControlNet



기존의 Diffusion 모델을 재학습

2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

2022.10

Prompt-to-Prompt

2022.12

InstructPix2Pix

ControlNet

InstructPix2Pix

"서부 영화로 만들어 줘"

대화형 수정 AI를 만들 수 있다

2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

2022.10

Prompt-to-Prompt

2022.12

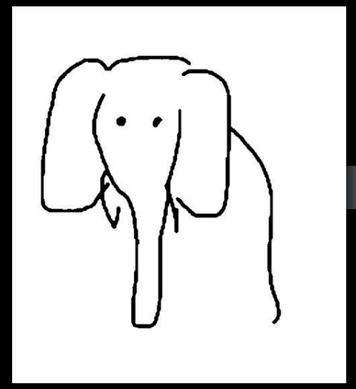
InstructPix2Pix

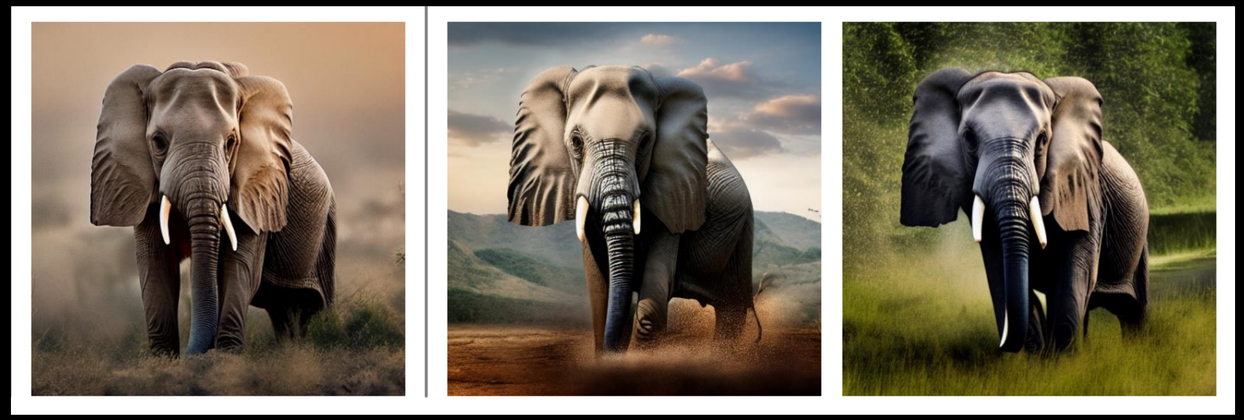
2022.02

ControlNet

ControlNet

Adding Conditional Control to Text-to-Image Diffusion Models





2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

2022.10

Prompt-to-Prompt

2022.12

InstructPix2Pix

2022.02

ControlNet

기존의 방식과 다른점?

텍스트가 아닌 **이미지**로

생성과 **수정**을 할 수 있는 방식

2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

xformers

2022.10

Prompt-to-Prompt

2022.12

InstructPix2Pix

2022.02

ControlNet

고작 6개월 만에 일어난 일

학습 비용

DreamBooth

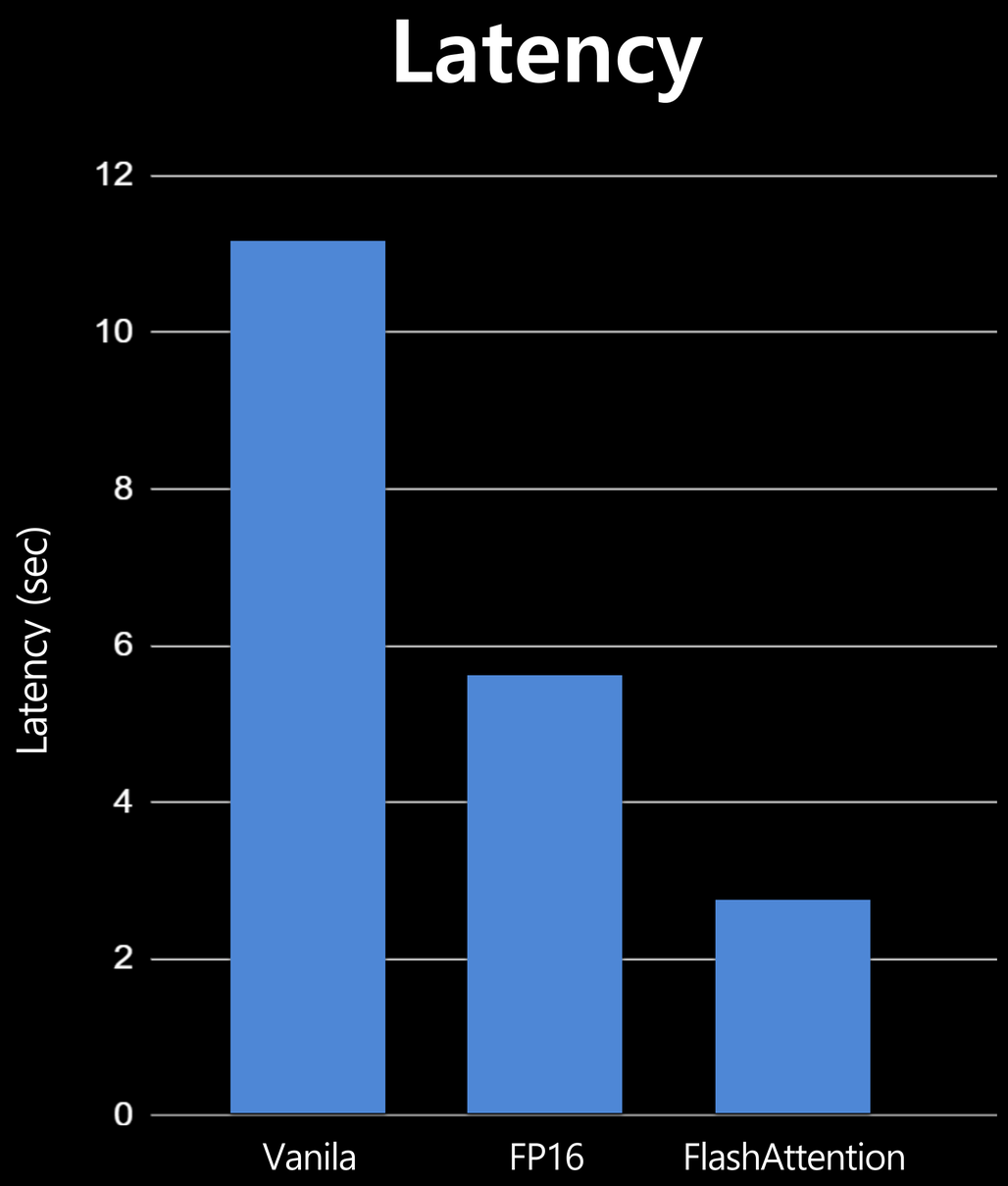
LoRA

이미지 수정

Prompt-to-Prompt

InstructPixPix

ControlNet



값비싼 Diffusion model를

저비용

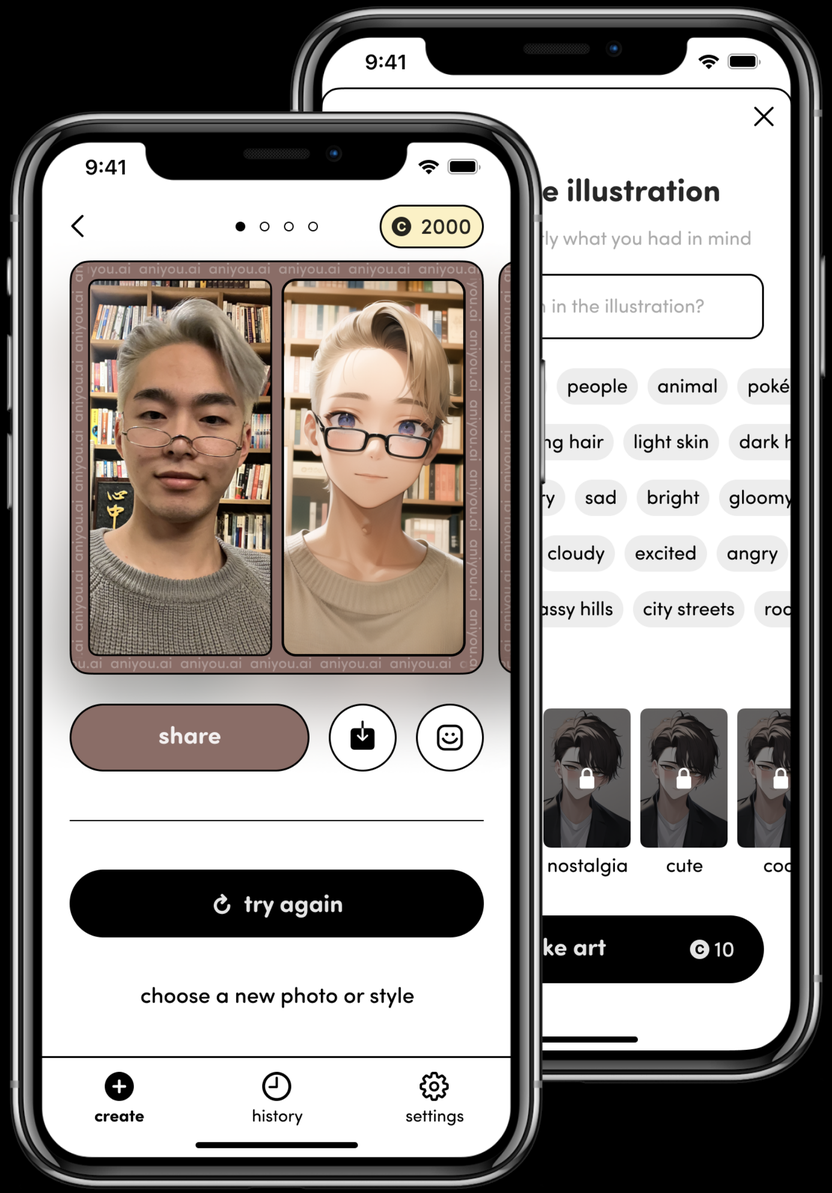
밭드는 MLOps

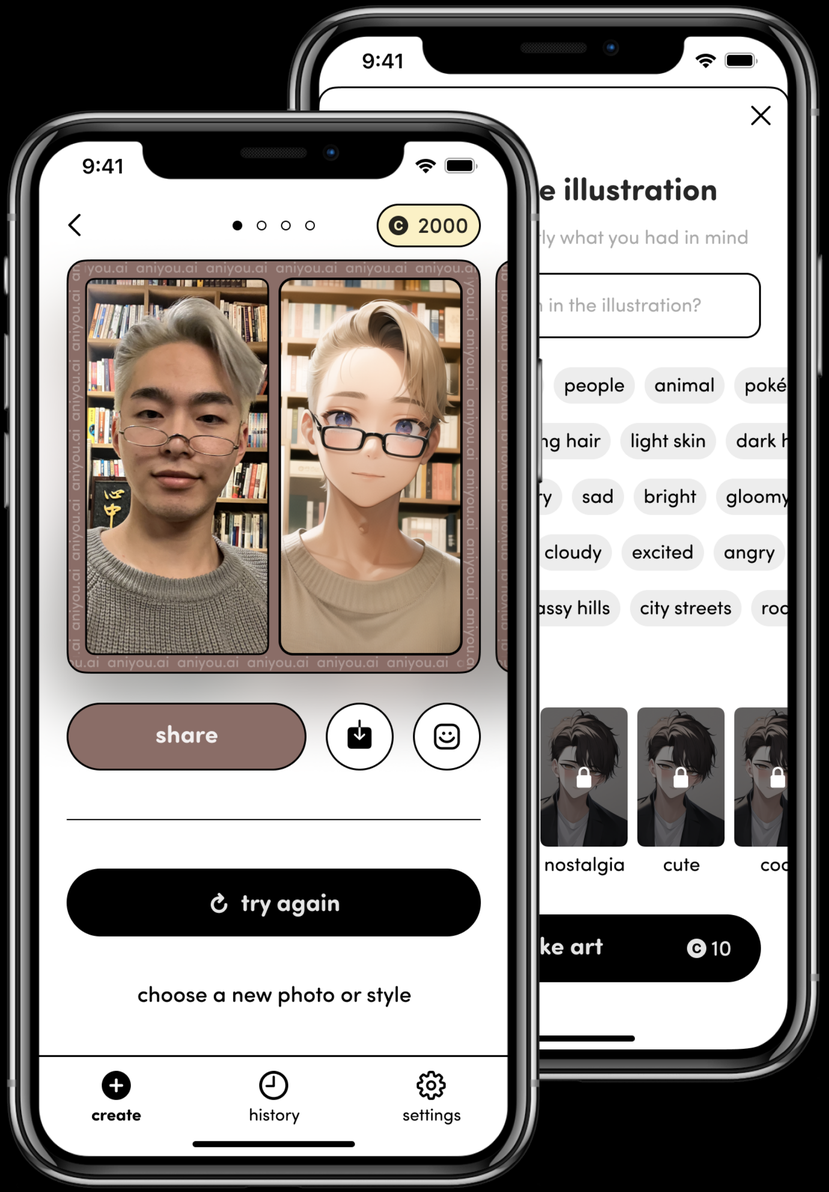
1. Diffusion은 무엇이 다른가?

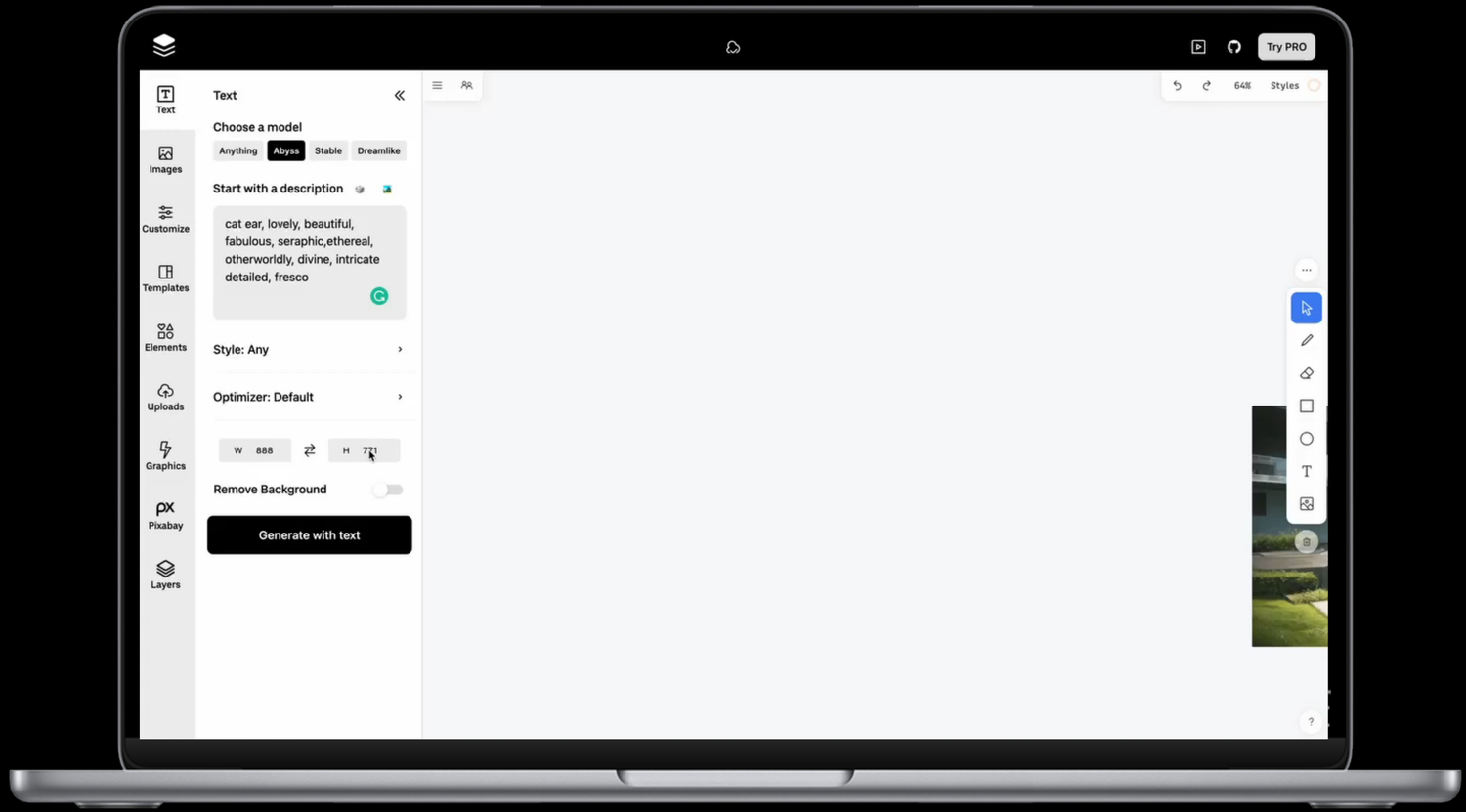
비용

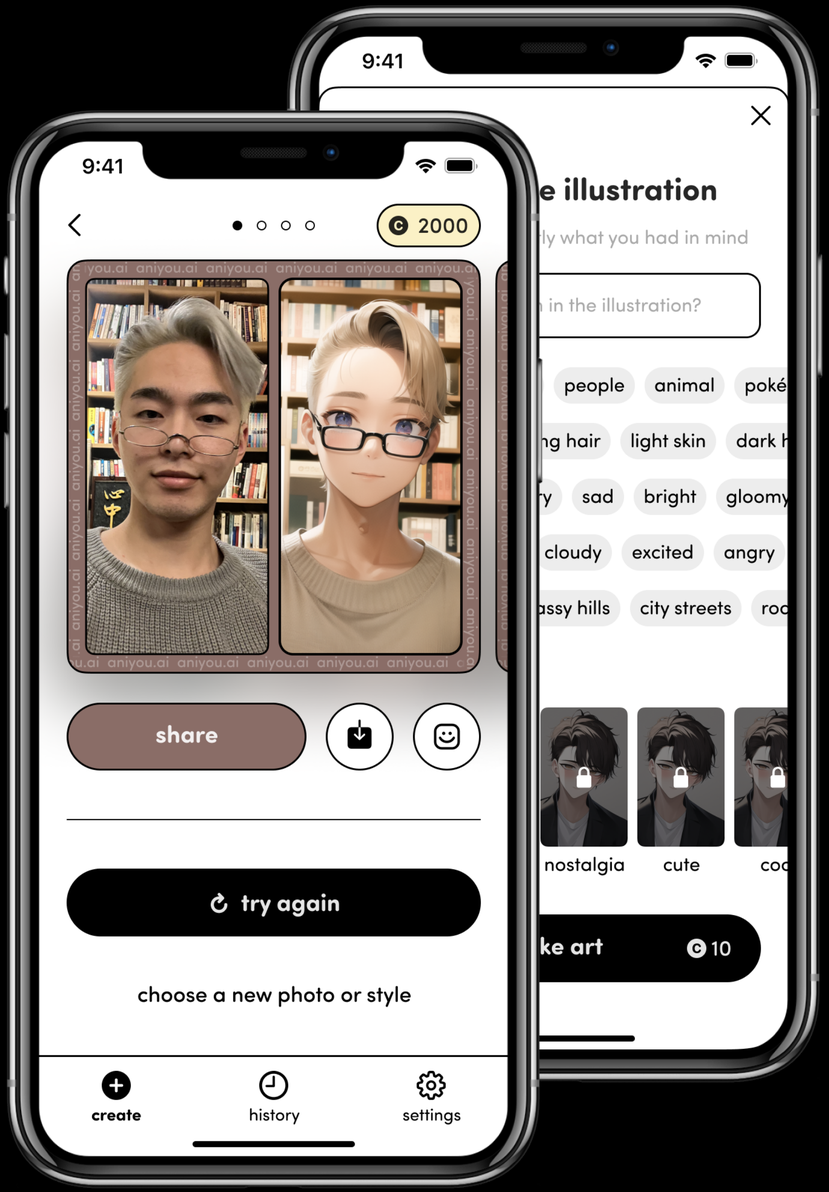
서비스 형태에 따라 다르다

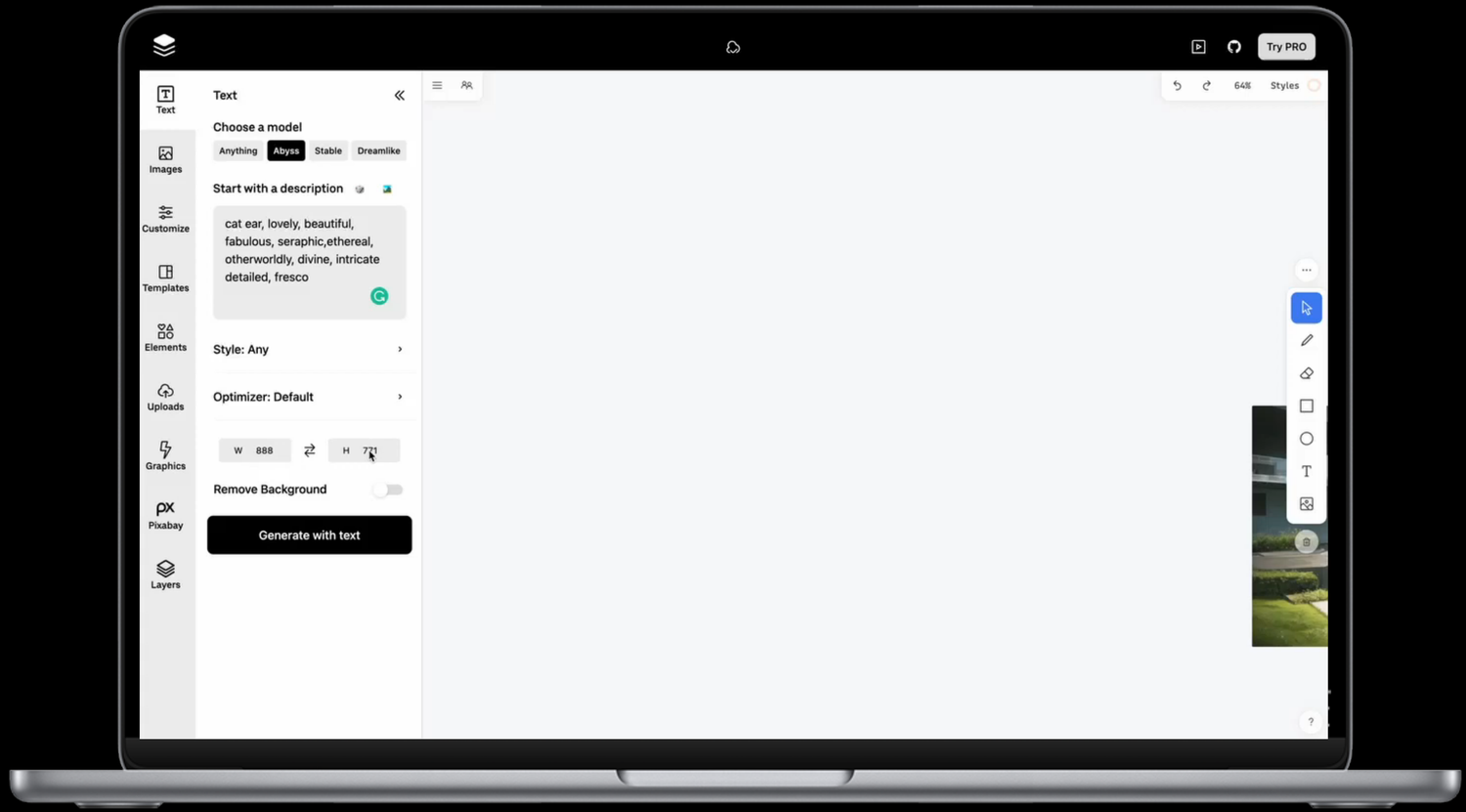
저희가 만들었던 서비스


9:41
2000
share
try again
choose a new photo or style
create
history
settings
9:41
e illustration
ly what you had in mind
in the illustration?
people
animal
poké
ng hair
light skin
dark h
ry
sad
bright
gloomy
cloudy
excited
angry
assy hills
city streets
roc
nostalgia
cute
coc
ke art
10


9:41
2000
share
try again
choose a new photo or style
create
history
settings
illustration
what you had in mind
in the illustration?
people
animal
light skin
sad
bright
cloudy
excited
angry
city streets
nostalgia
cute
10

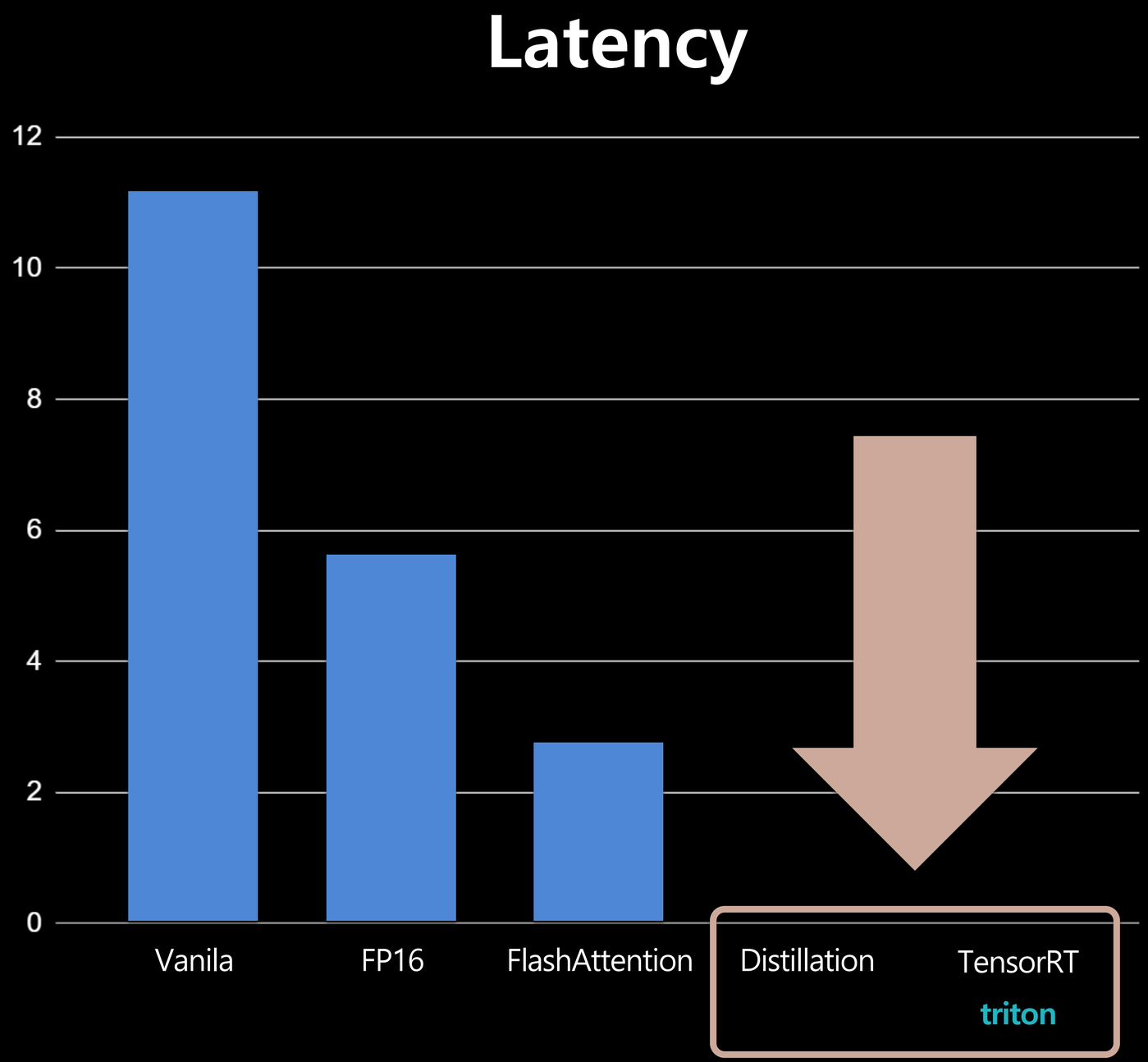

Try PRO
Text
Images
Customize
Templates
Elements
Uploads
Graphics
Pixabay
Layers
Choose a model
Anything
Abyss
Stable
Dreamlike
Start with a description
cat ear, lovely, beautiful, fabulous, seraphic,ethereal, otherworldly, divine, intricate detailed, fresco
Style: Any
Optimizer: Default
W 888
H 771
Remove Background
Generate with text
64%
Styles


9:41
2000
share
try again
choose a new photo or style
create
history
settings
illustration
what you had in mind
in the illustration?
people
animal
light skin
sad
bright
cloudy
excited
angry
city streets
nostalgia
cute
10


Try PRO
Text
Images
Customize
Templates
Elements
Uploads
Graphics
Pixabay
Layers
Choose a model
Anything
Abyss
Stable
Dreamlike
Start with a description
cat ear, lovely, beautiful, fabulous, seraphic,ethereal, otherworldly, divine, intricate detailed, fresco
Style: Any
Optimizer: Default
W 888
H 771
Remove Background
Generate with text
64%
Styles

Latency와 Scalability가 중요

Latency


Latency
12
10
8
6
4
2
0
Vanila
FP16
FlashAttention
Distillation
TensorRT
triton

1. Distillation by 이동익

2. TensorRT &  triton

1. Distillation by 이동익

2. TensorRT &  triton

@bryandlee

1. Distillation by 이동익

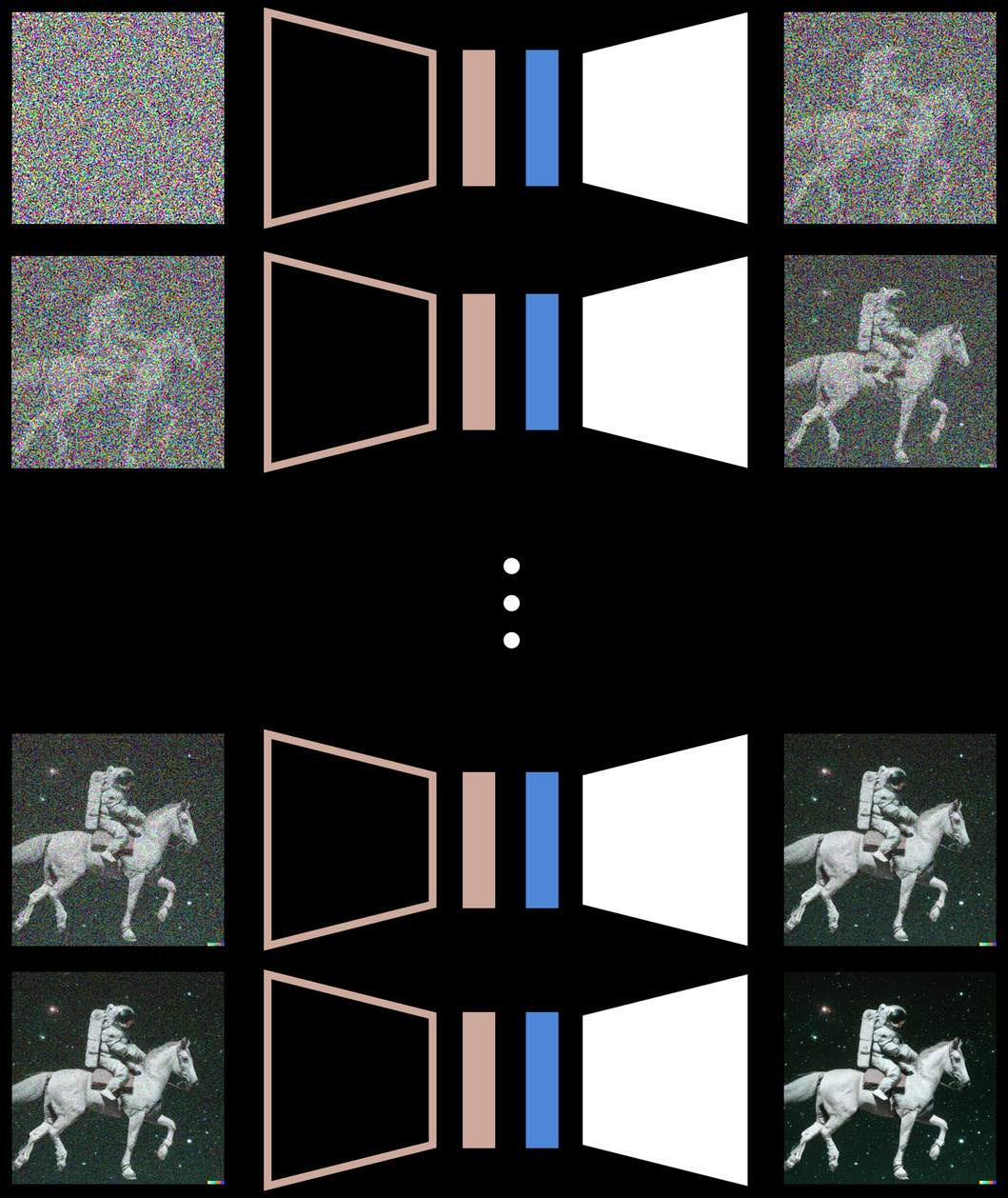
https://github.com/bryandlee/malnyun_faces

@bryandlee

1. Distillation by 이동익

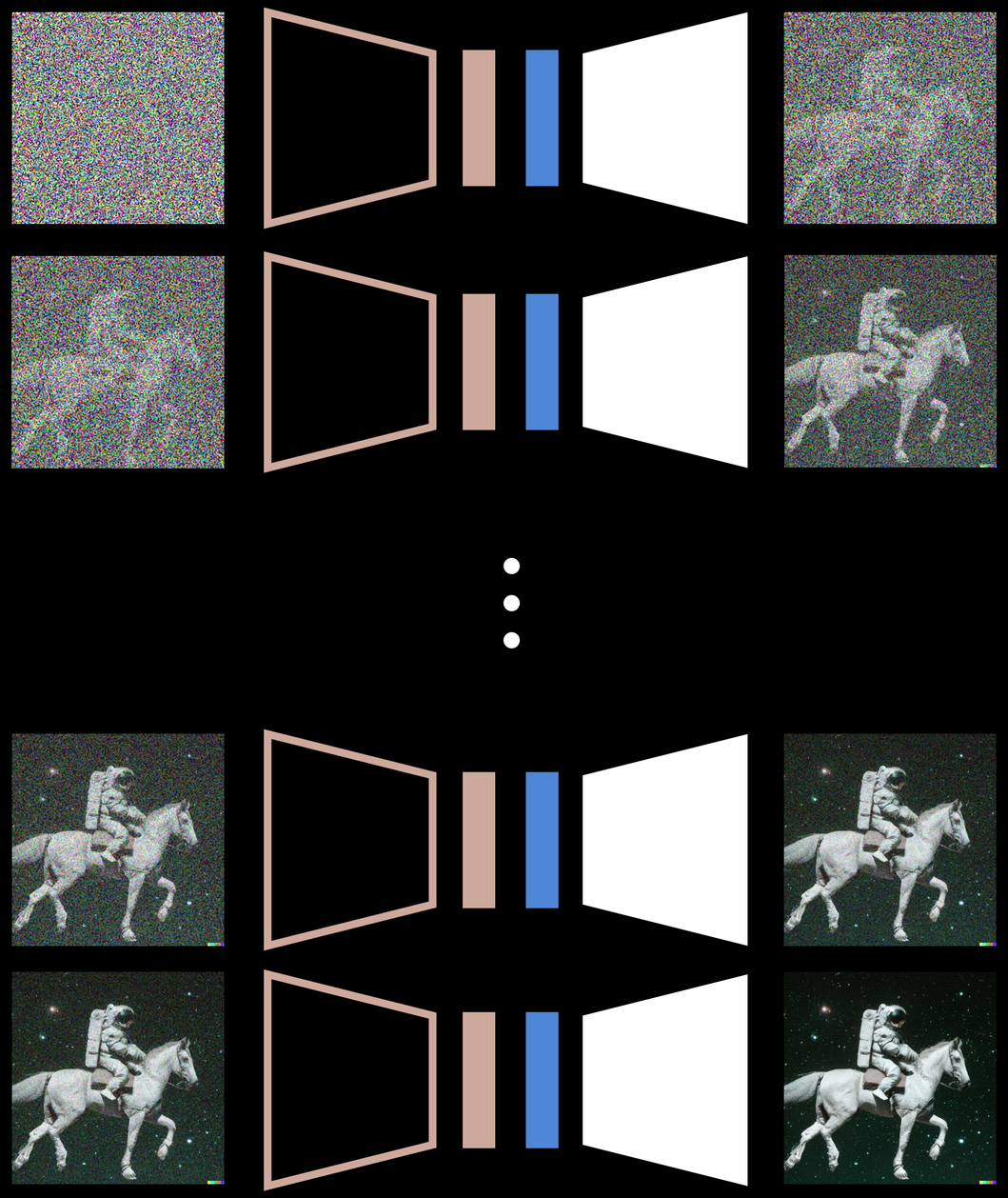
https://github.com/bryandlee/malnyun_faces

https://github.com/bryandlee/DeepStudio



50 step

V100 기준 10초



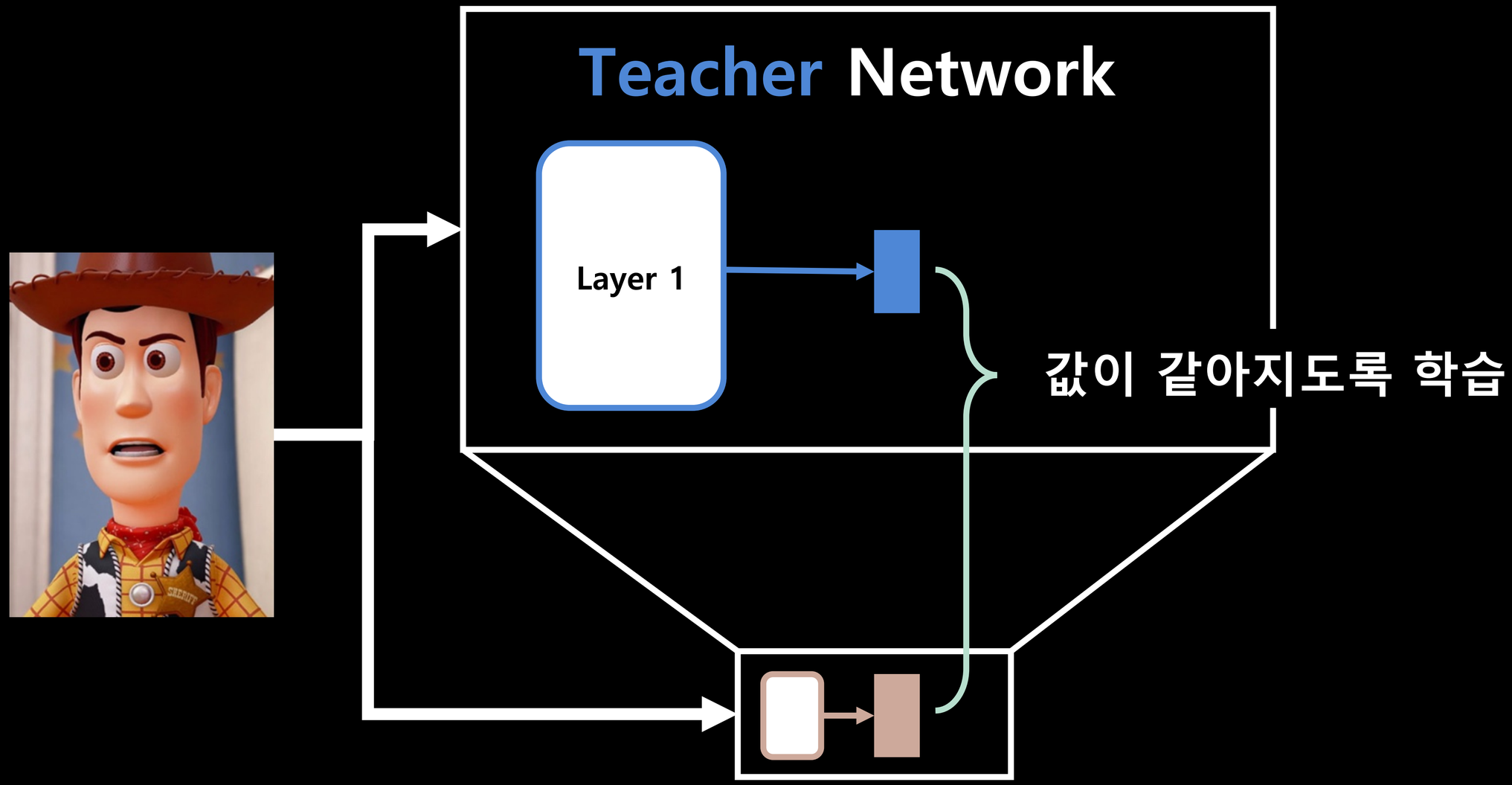
추론 횟수를
줄이려면?

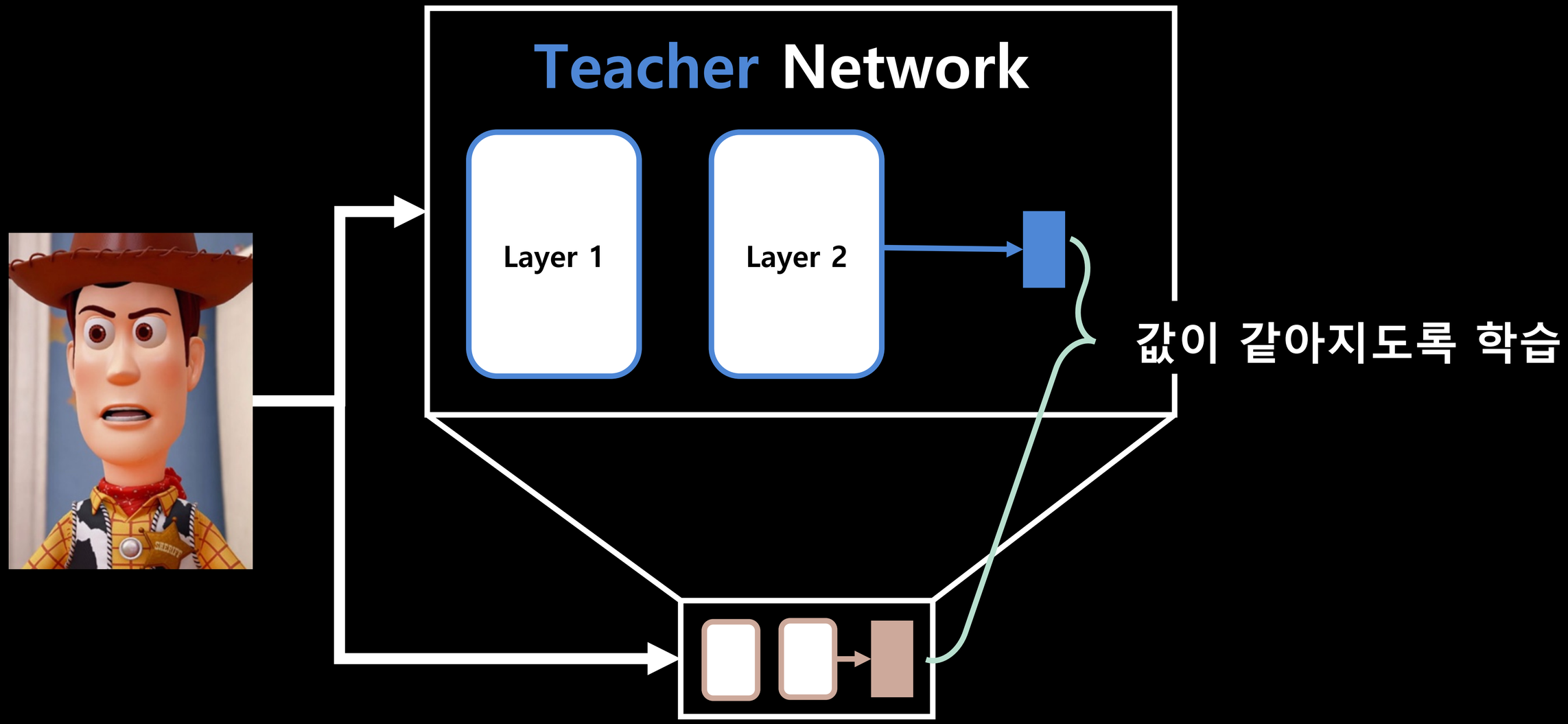
Knowledge Distillation

Teacher Network

Teacher Network

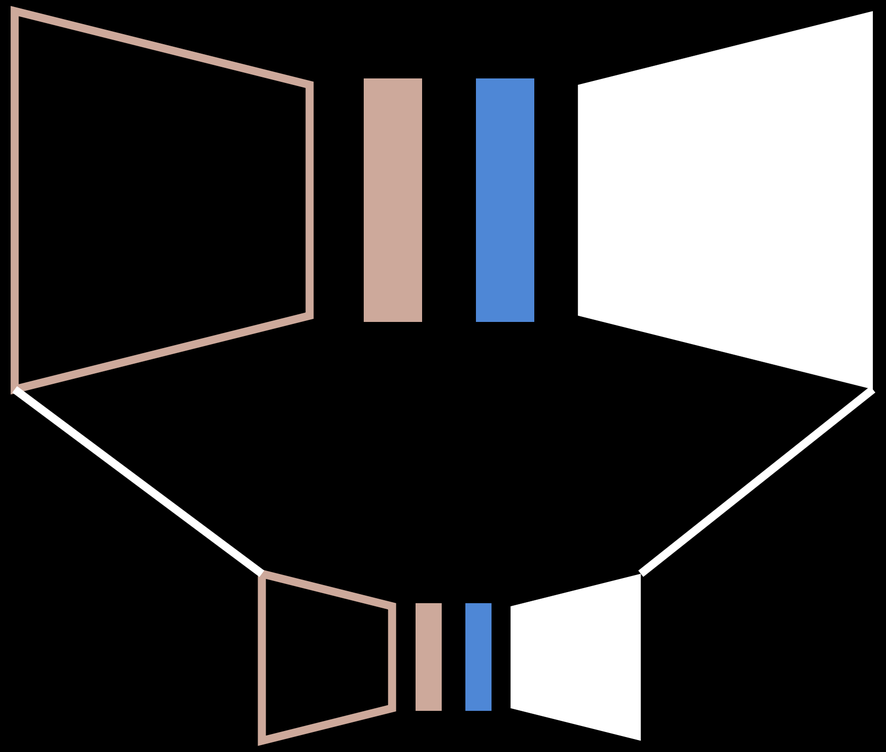
Student Network


Teacher Network
Layer 1
값이 같아지도록 학습
Student Network


Teacher Network
Layer 1
Layer 2
값이 같아지도록 학습
Student Network

Knowledge Distillation

= 모델 다이어트



Diffusion 모델도

모델 크기를 줄일 수 있지만

On Distillation of Guided Diffusion Models

Chenlin Meng*
Stanford University
chenlin@cs.stanford.edu

Robin Rombach
Stability AI & LMU Munich
robin@stability.ai

Ruiqi Gao
Google Research, Brain Team
ruiqig@google.com

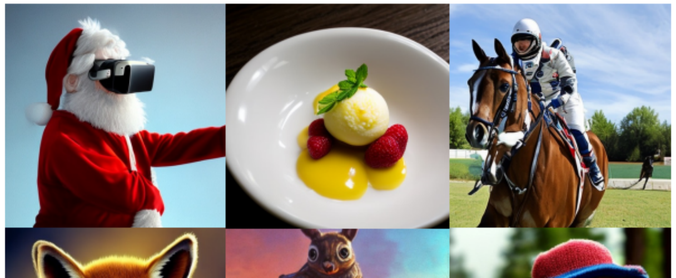
Diederik P. Kingma
Google Research, Brain Team
durk@google.com

Stefano Ermon
Stanford University
ermon@cs.stanford.edu

Jonathan Ho
Google Research, Brain Team
jonathanho@google.com

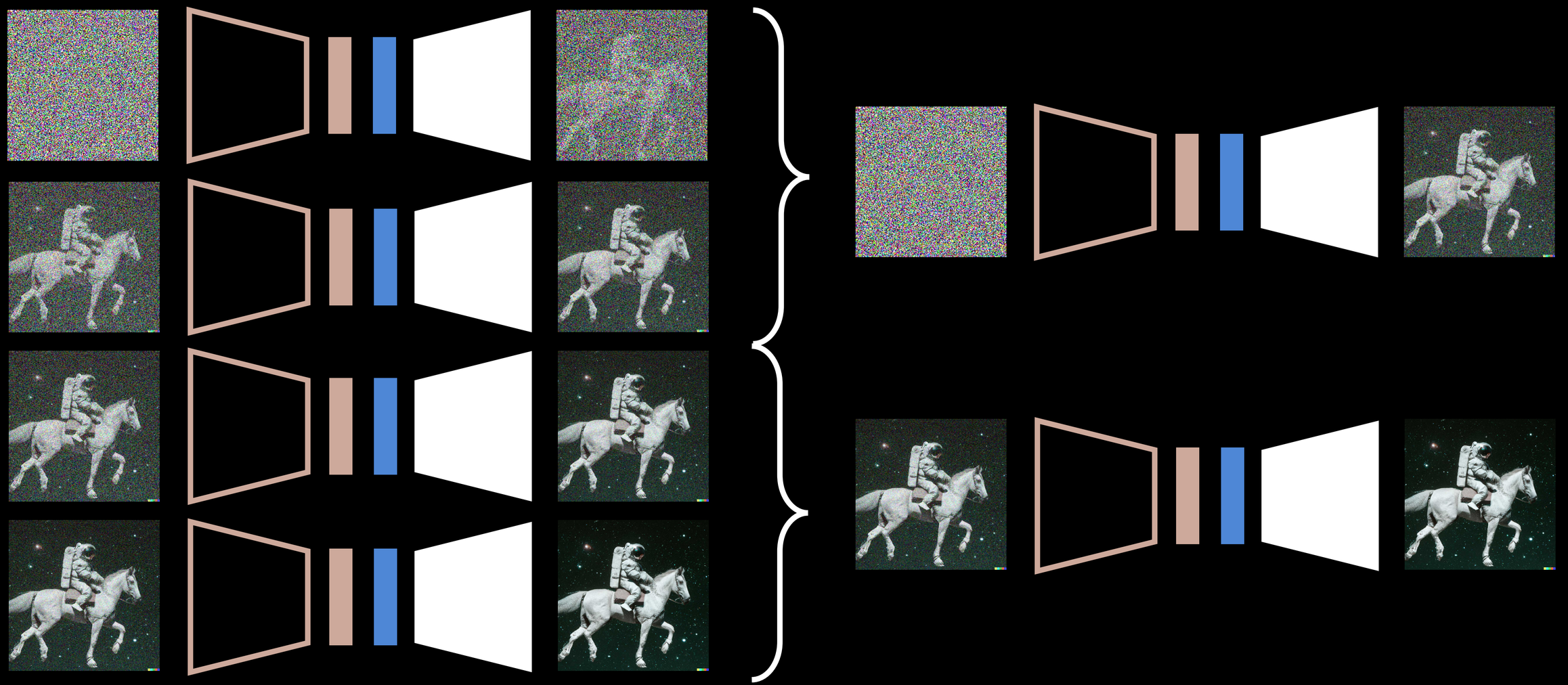
Tim Salimans
Google Research, Brain Team
salimans@google.com

Text-guided generation (1 step)

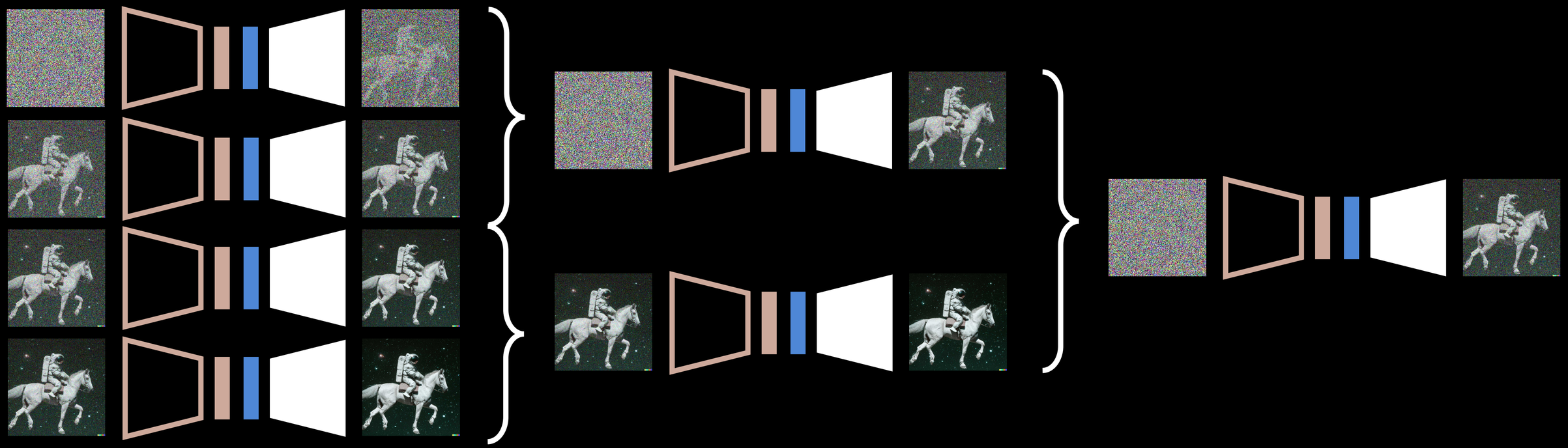


Text-guided generation (2 steps)

Class-conditional generation (1 step)

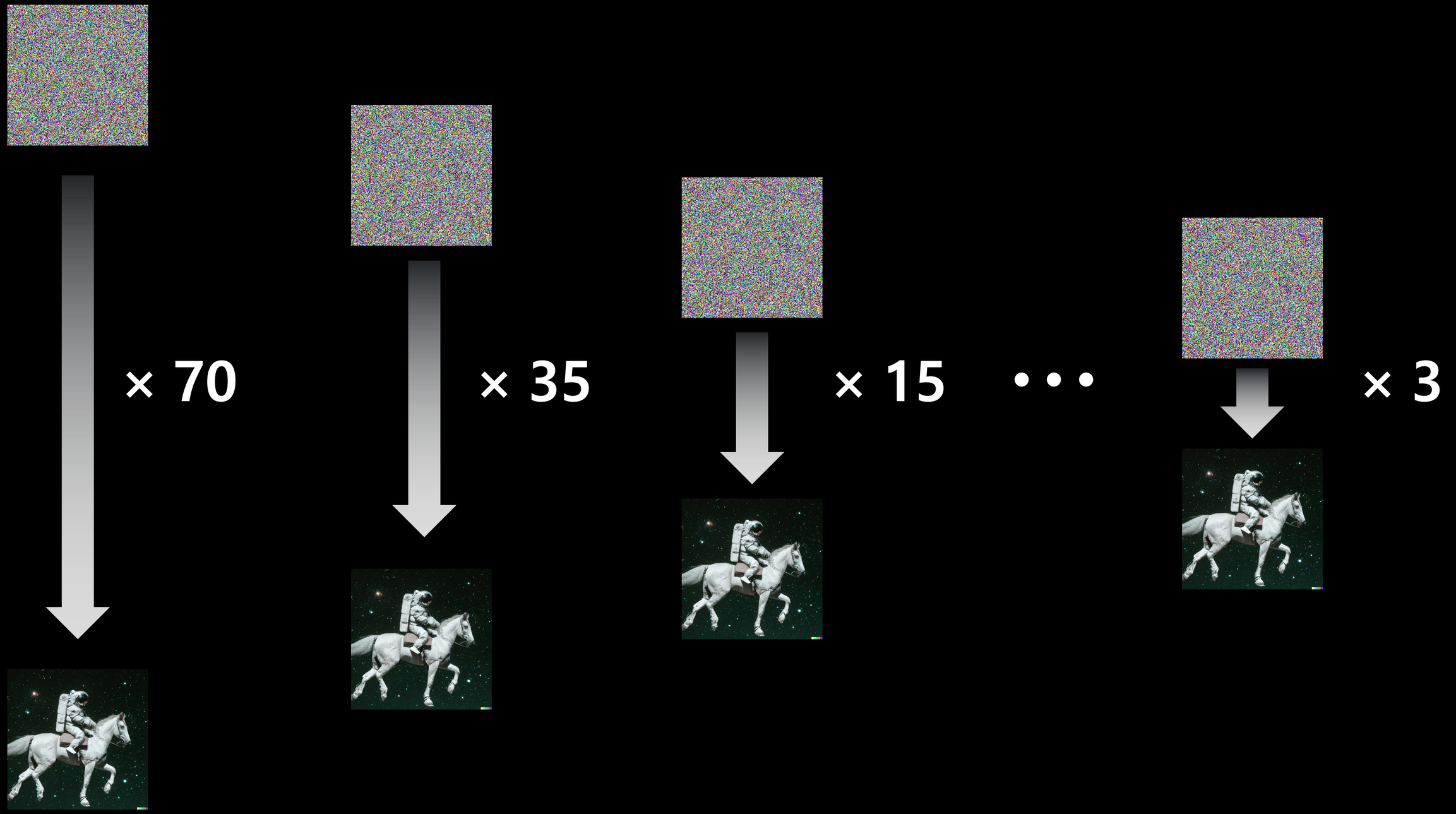


Distillation을 할 때 마다 절반씩 줄어든다



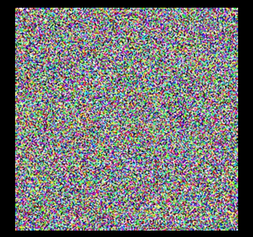
Distillation of diffusion step

= 스텝 다이어트


× 70
× 35
× 15
• • •
× 3

× 35

10배 빨라짐



× 3

실험 결과

1. Image2Image 모델

2. 기존 모델이 학습한 데이터 셋에서 학습

3. 1번의 distillation은 평균 1.5일 정도 걸림

4. 전체 학습 기간은 6일

평범한 Diffusion

× 35

Distillation 없이 3번

Distillation 학습 후 3번

× 3

× 35

Distillation 없이 3번

t = 35

t = 25

t = 15

t = 0

Distillation 학습 후 3번

× 3

“일러스트화”

“인간화”

“꽃 추가”

"일러스트화"

"인간화"

"꽃 추가"

Free

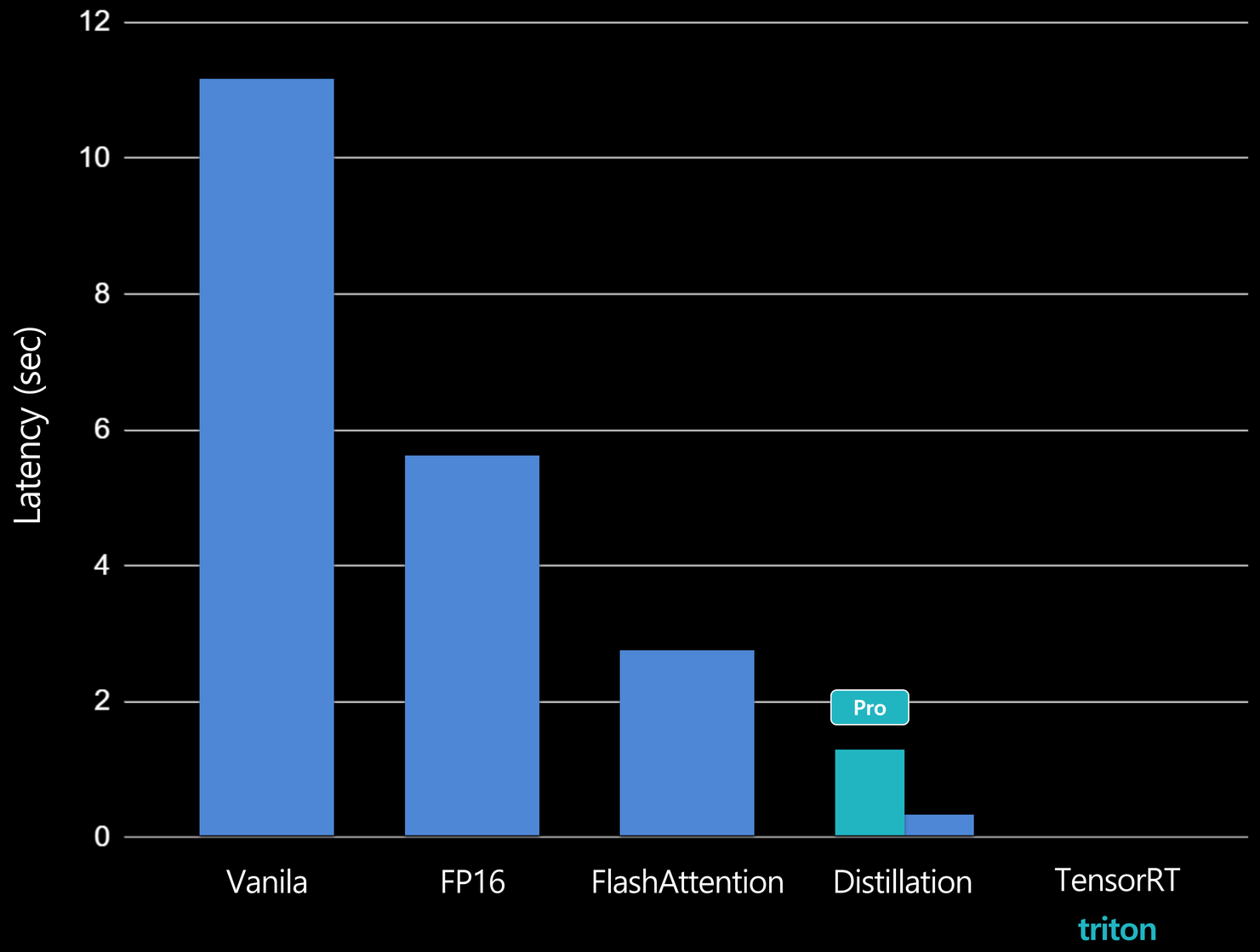
× 3

Pro

완성도를 중요시 하는 유저

× 35

Latency


12
10
8
6
4
2
0
Latency (sec)
Pro
Vanila
FP16
FlashAttention
Distillation
TensorRT
triton

1. Distillation by 이동익

2. TensorRT &  triton

1. Distillation by 이동익

2. TensorRT &  triton

2021.12 ~ 2022.08

Stable Diffusion

2022.08

DreamBooth & LoRA

2022.09

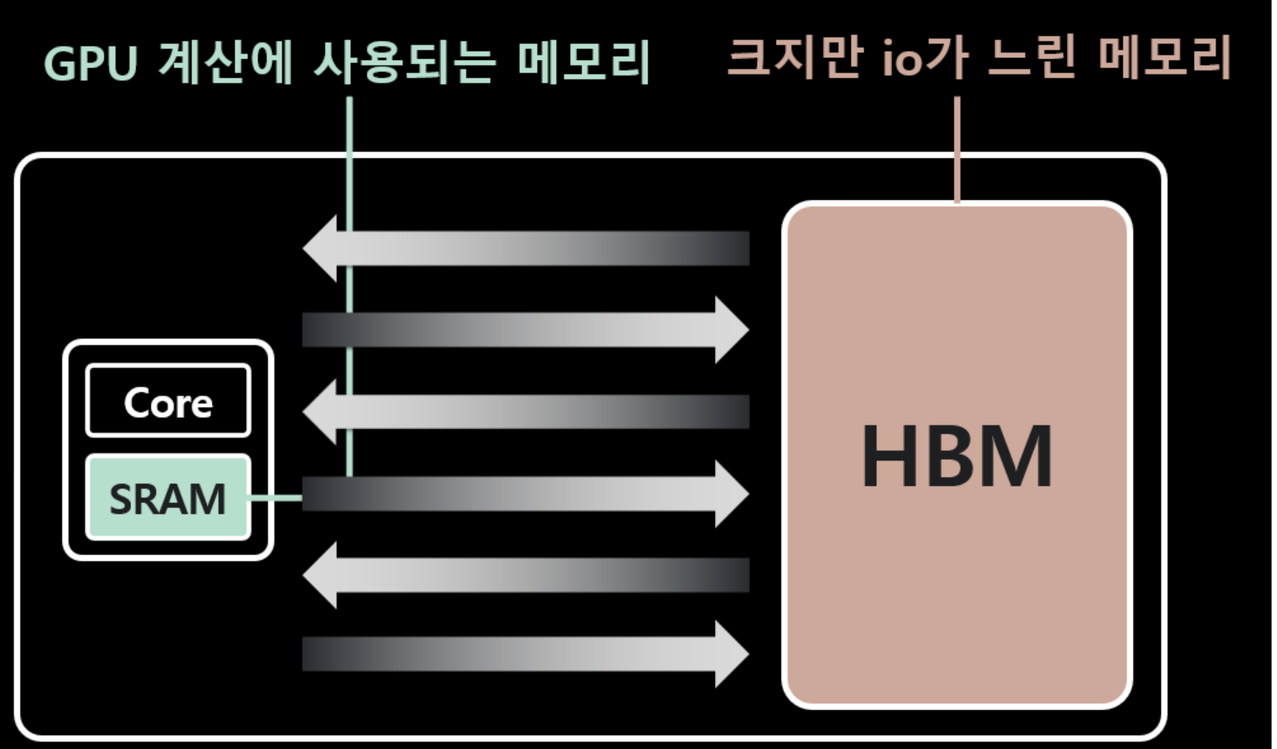
xformers

Prompt-to-Prompt

InstructPix2Pix

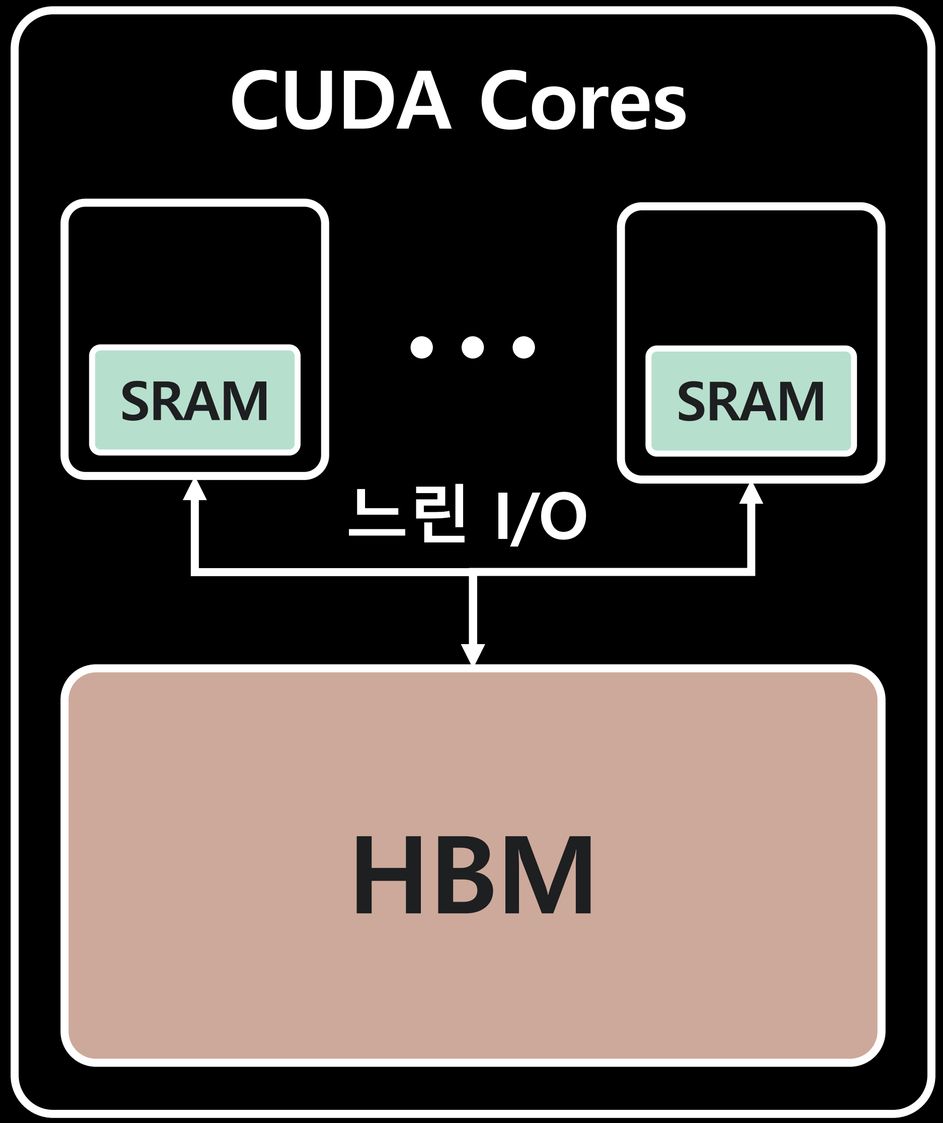
ControlNet

FlashAttention

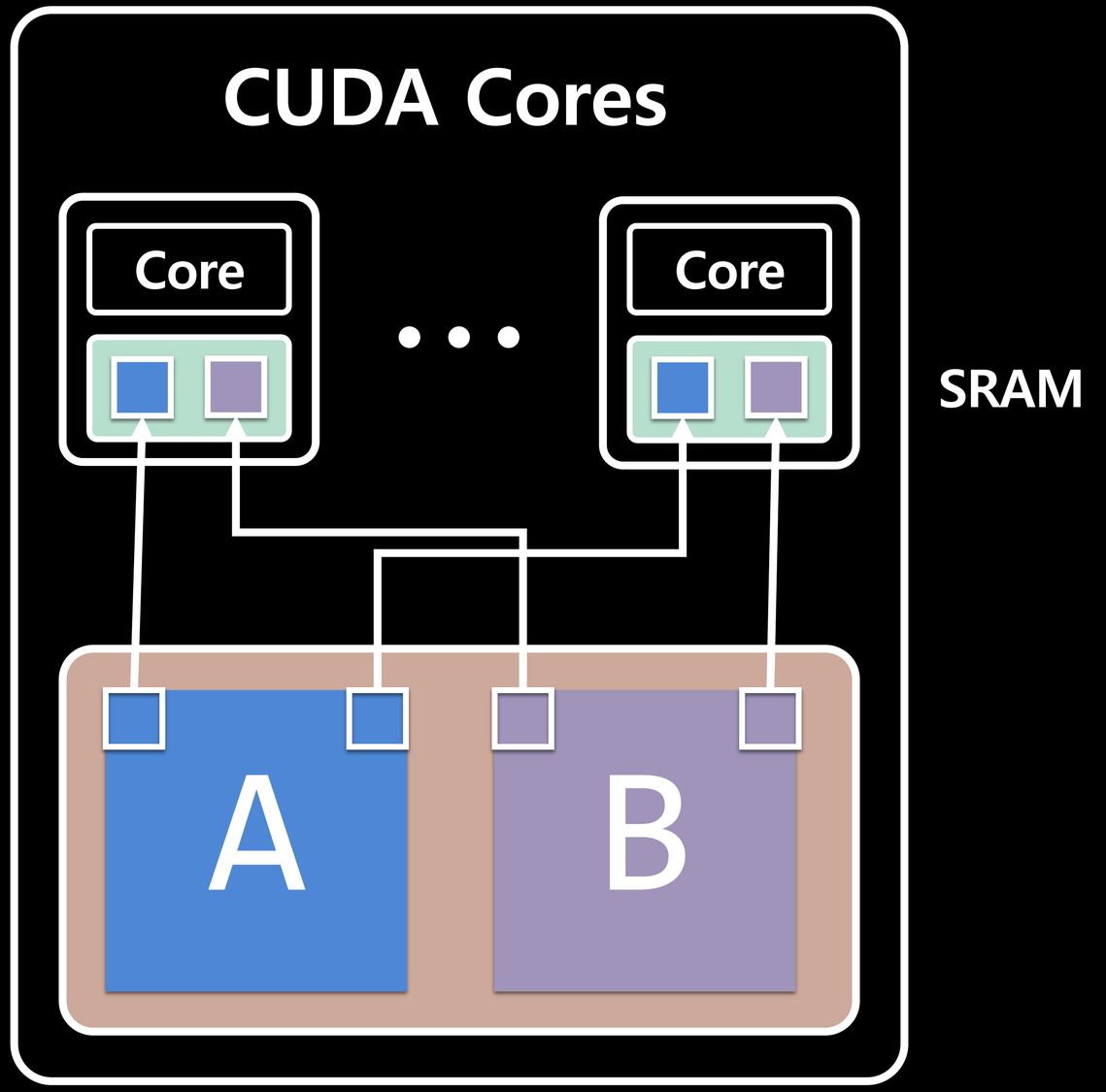


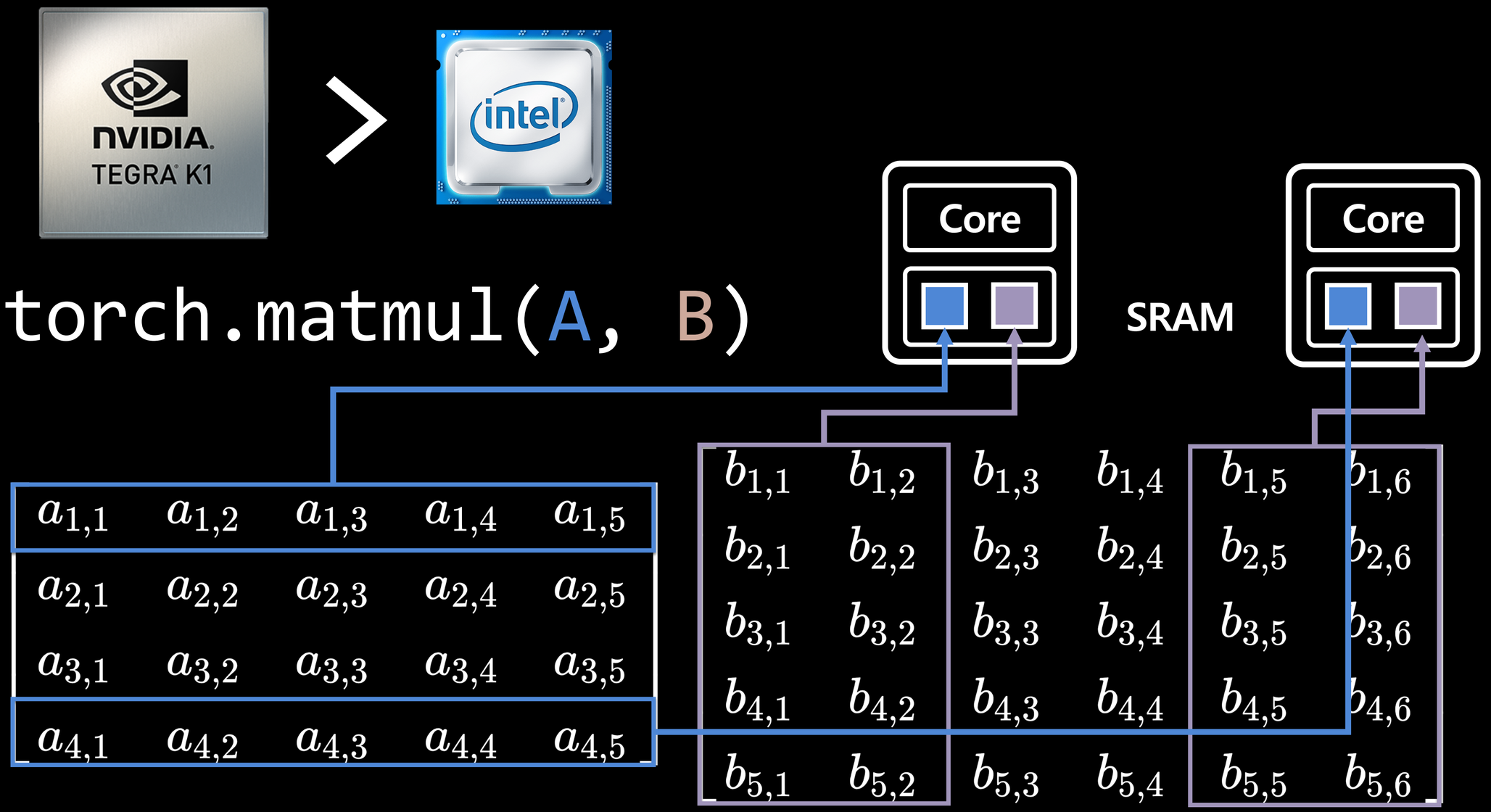
GPU 하드웨어 구조

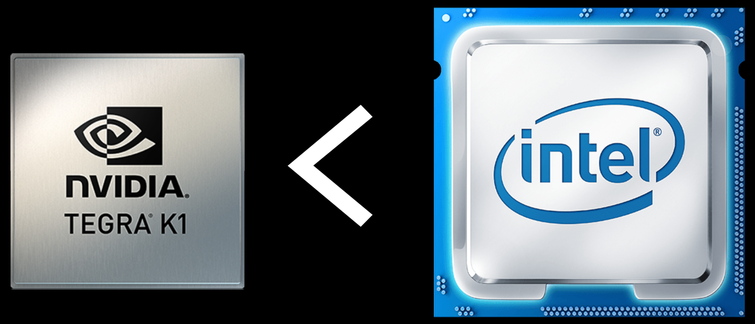

NVIDIA
TEGRA K1
RTX 4090


CUDA Cores
SRAM
SRAM
느린 I/O
HBM

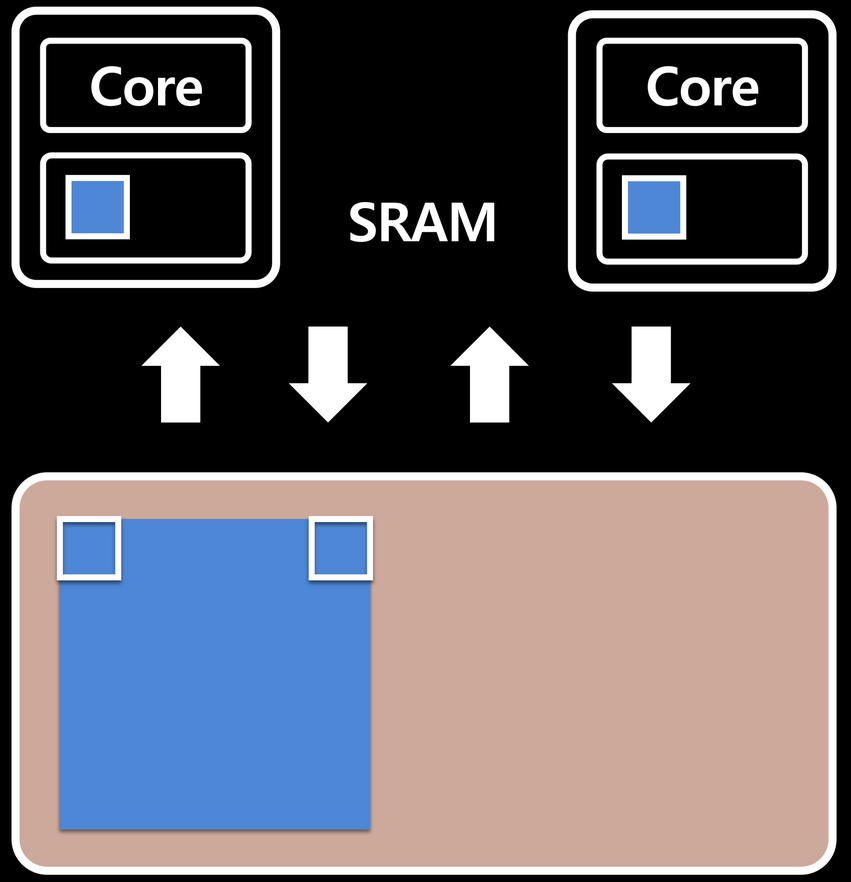
```
torch.matmul(A, B)
```

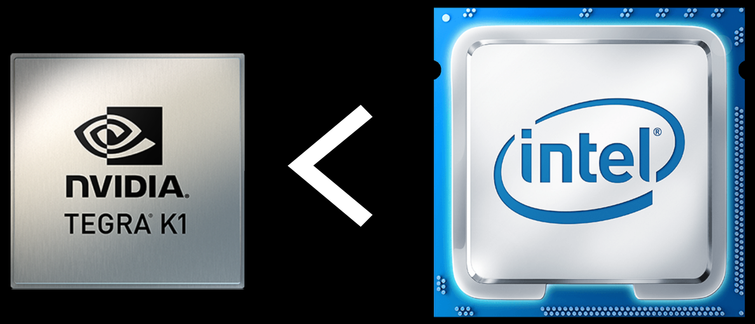



NVIDIA
TEGRA K1
intel
torch.matmul(A, B)
Core
SRAM
Core
$a_{1,1}$ $a_{1,2}$ $a_{1,3}$ $a_{1,4}$ $a_{1,5}$
$a_{2,1}$ $a_{2,2}$ $a_{2,3}$ $a_{2,4}$ $a_{2,5}$
$a_{3,1}$ $a_{3,2}$ $a_{3,3}$ $a_{3,4}$ $a_{3,5}$
$a_{4,1}$ $a_{4,2}$ $a_{4,3}$ $a_{4,4}$ $a_{4,5}$
$b_{1,1}$ $b_{1,2}$ $b_{1,3}$ $b_{1,4}$ $b_{1,5}$ $b_{1,6}$
$b_{2,1}$ $b_{2,2}$ $b_{2,3}$ $b_{2,4}$ $b_{2,5}$ $b_{2,6}$
$b_{3,1}$ $b_{3,2}$ $b_{3,3}$ $b_{3,4}$ $b_{3,5}$ $b_{3,6}$
$b_{4,1}$ $b_{4,2}$ $b_{4,3}$ $b_{4,4}$ $b_{4,5}$ $b_{4,6}$
$b_{5,1}$ $b_{5,2}$ $b_{5,3}$ $b_{5,4}$ $b_{5,5}$ $b_{5,6}$

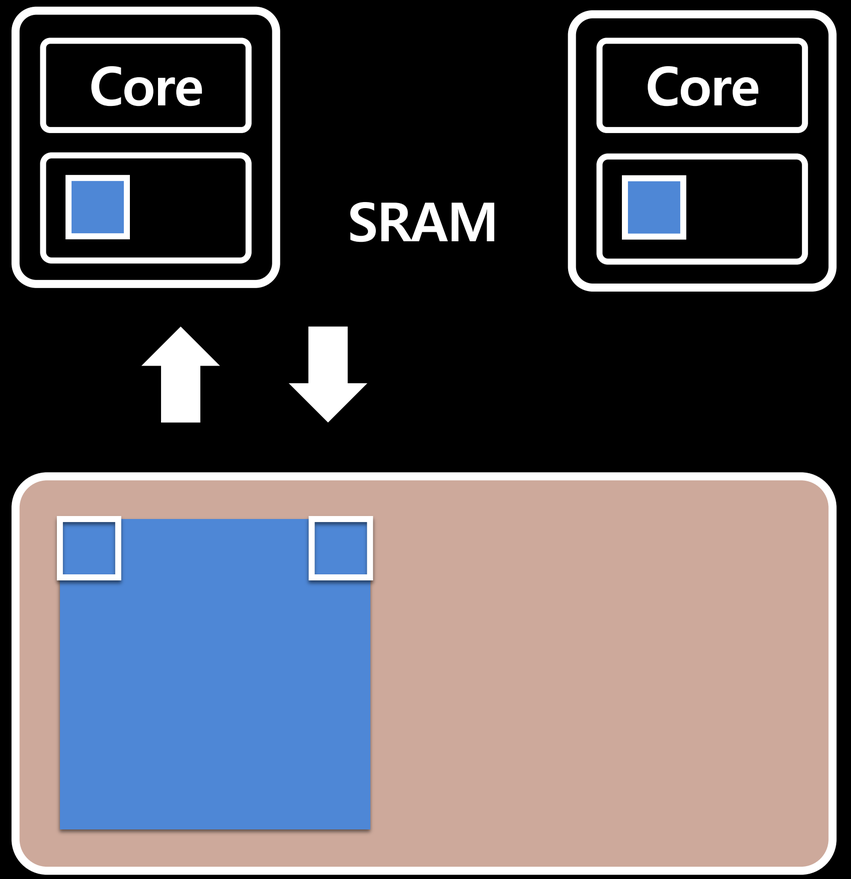


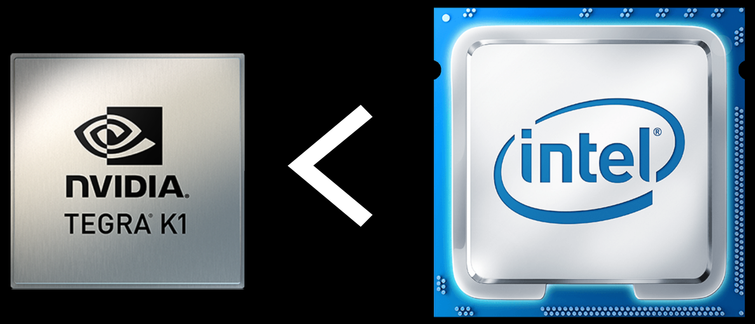

```
A.cuda()
A = A.cos()
A = A.cos()
```



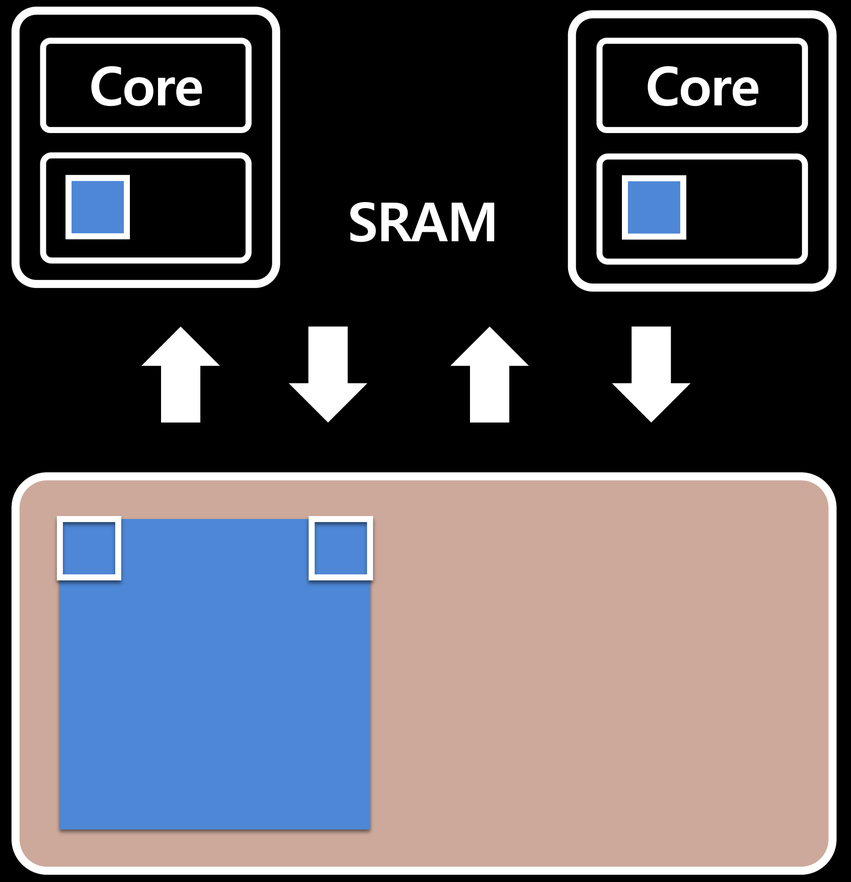


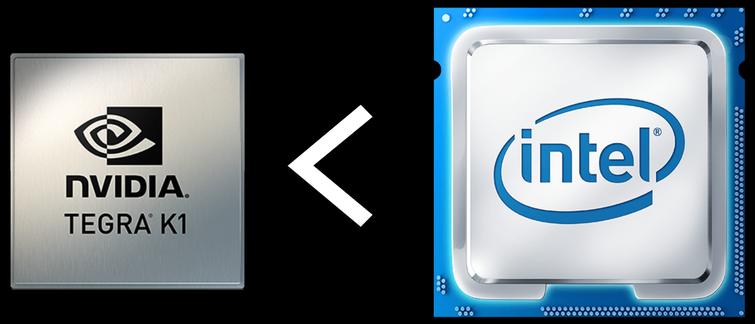

```
A = A.cos()
```



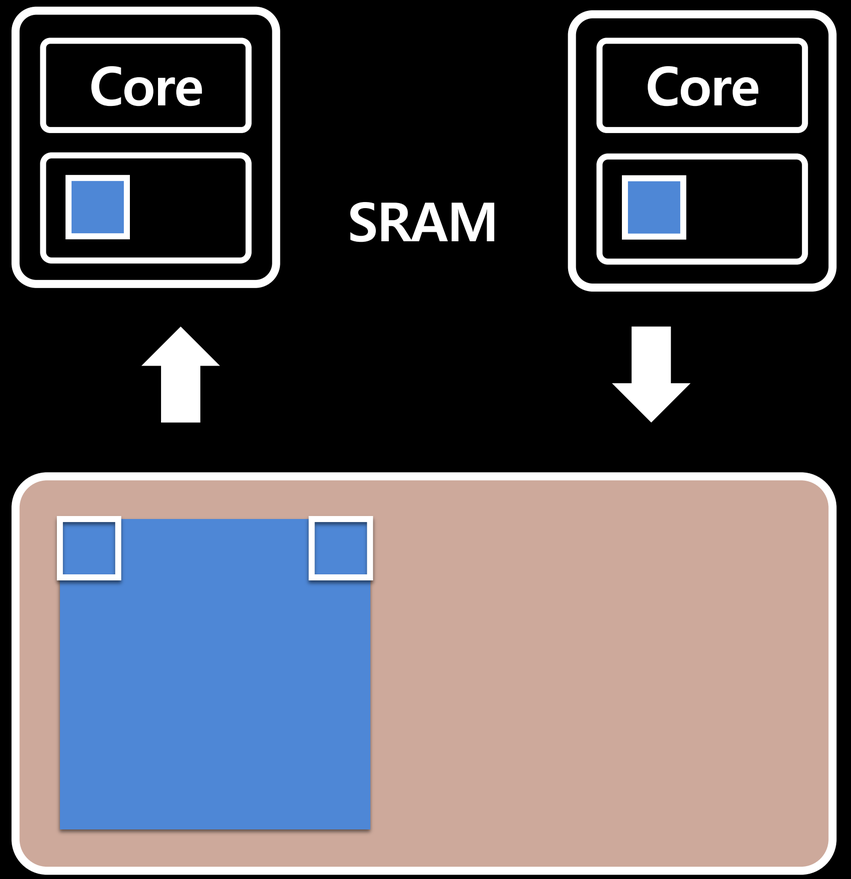



```
A = A.cos()
A = A.cos()
```



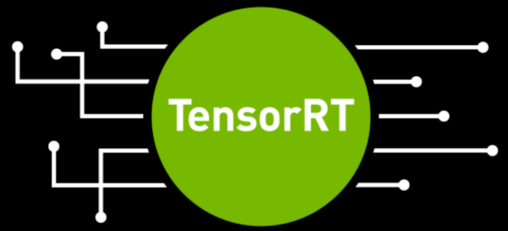



```
A = A.cos().cos()
```

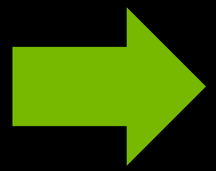


```
A.cos().cos()
```

```
swish(A)

gelu(A)

GroupNorm(A)

LayerNorm(A)

...
```

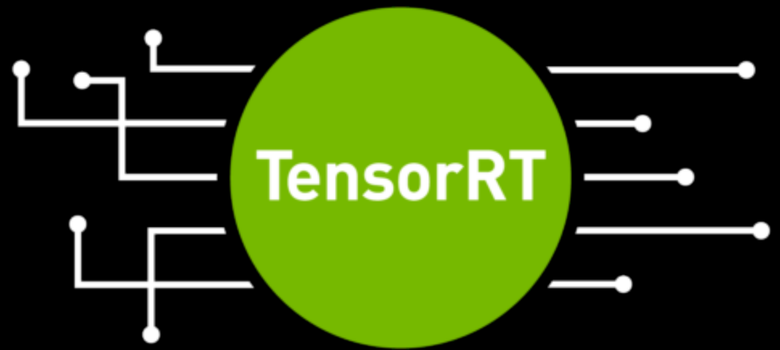



```
A.cos().cos()
```



```
A.cos_cos()
```

NVIDIA



를 쓰는 방법


```
model = Model()
```


```
model = Model()

onnx = Onnx.export(model)
```


```
model = Model()

onnx = Onnx.export(model)

new_model = trt.compile(onnx)
```


```
model = Model()

onnx = Onnx.export(model)

new_model = trt.compile(onnx)

y = new_model(x)
```

```
new_model = trt.compile(onnx)
```

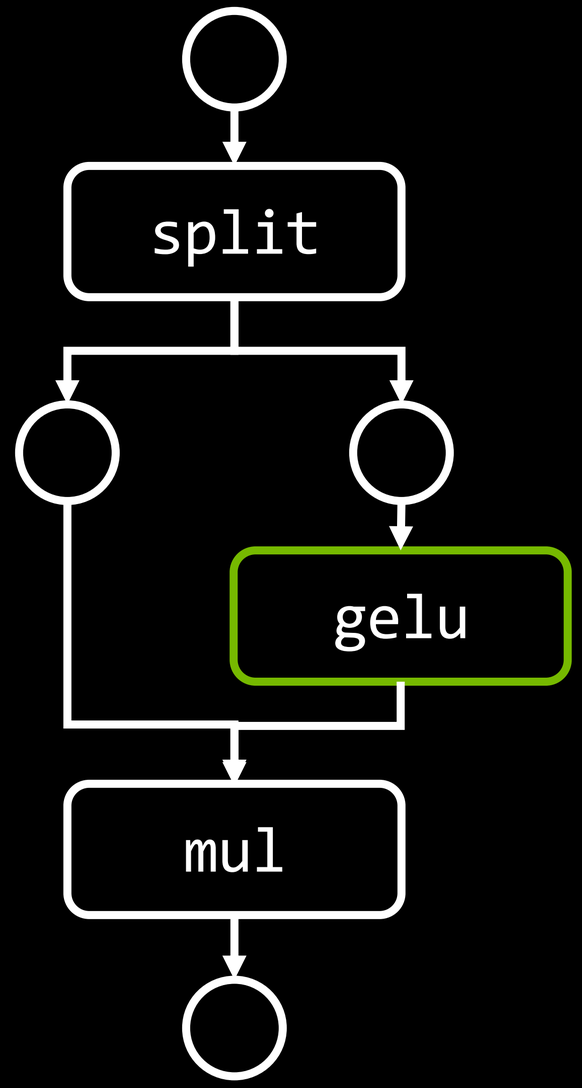
이것만 쓰면 되는 건가?

Nope.

Static graph 만으론

모든걸 알 순 없음

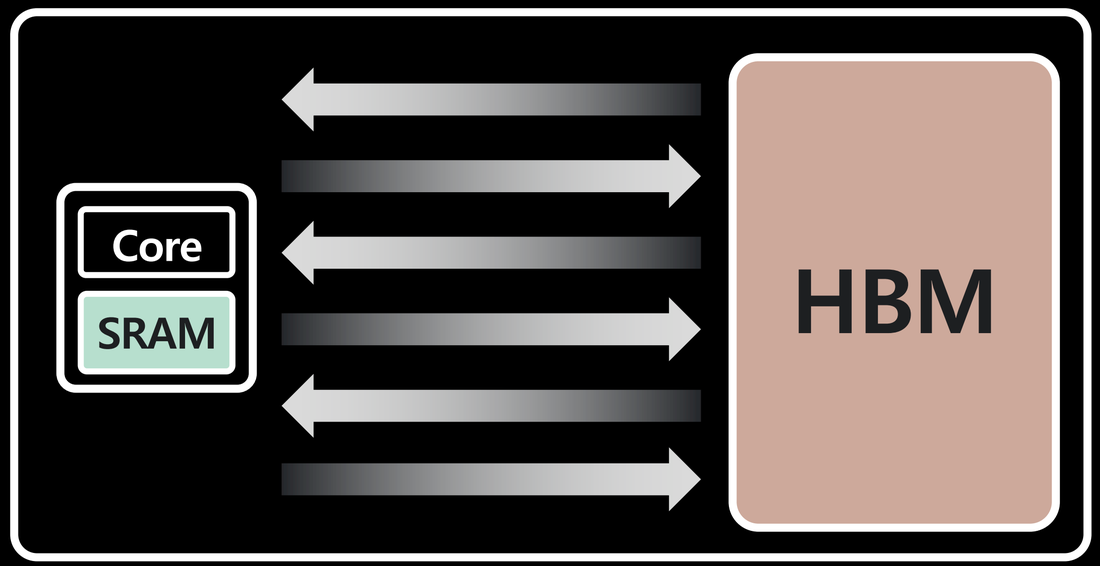
```python
class GEGLU(nn.Module):
  def __init__(self, dim_in, dim_out):
    super().__init__()
    self.proj = nn.Linear(dim_in, dim_out * 2)

  def forward(self, x):
    x, gate = self.proj(x).chunk(2, dim=-1)
    return x * F.gelu(gate)
```

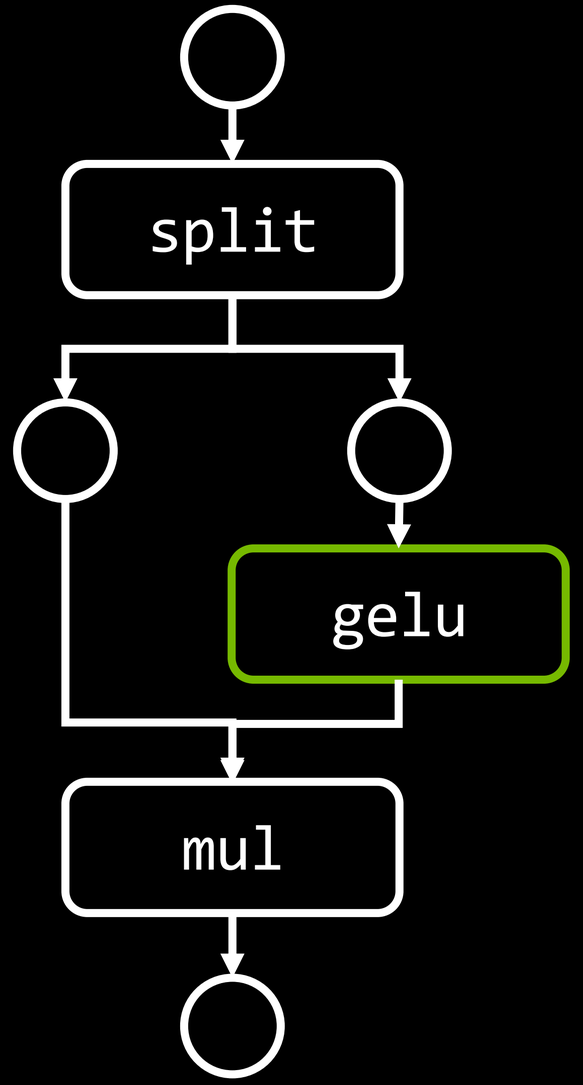


https://github.com/CompVis/latent-diffusion/blob/main/ldm/modules/attention.py#L37-L44

Unary operation (단항 연산자)

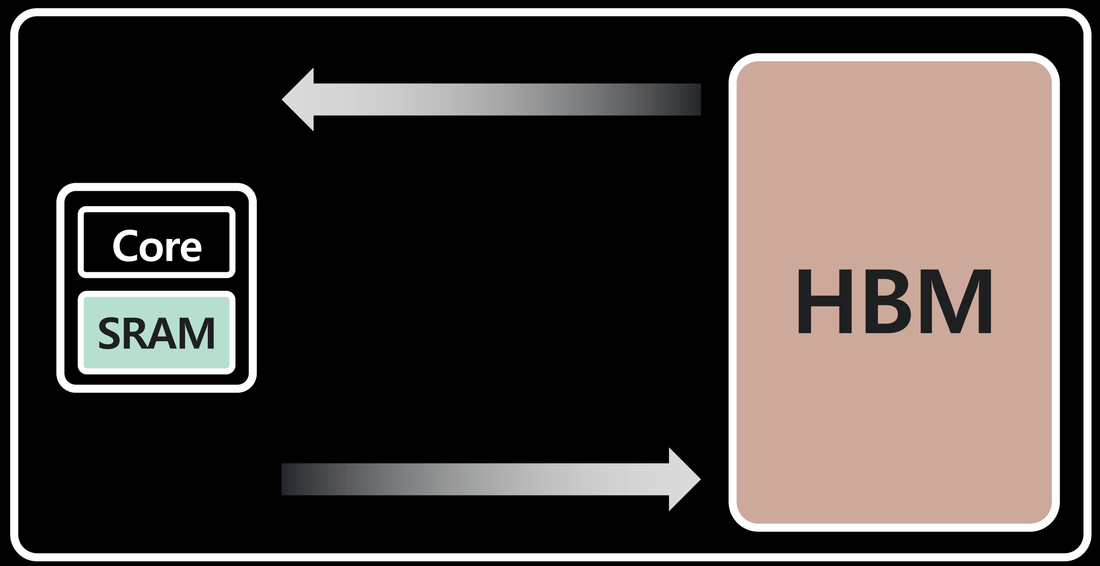
$$GELU(x) \cong 0.5x(1 + \tanh\left[\sqrt{\frac{2}{\pi}}(x + 0.4x^3)\right]$$

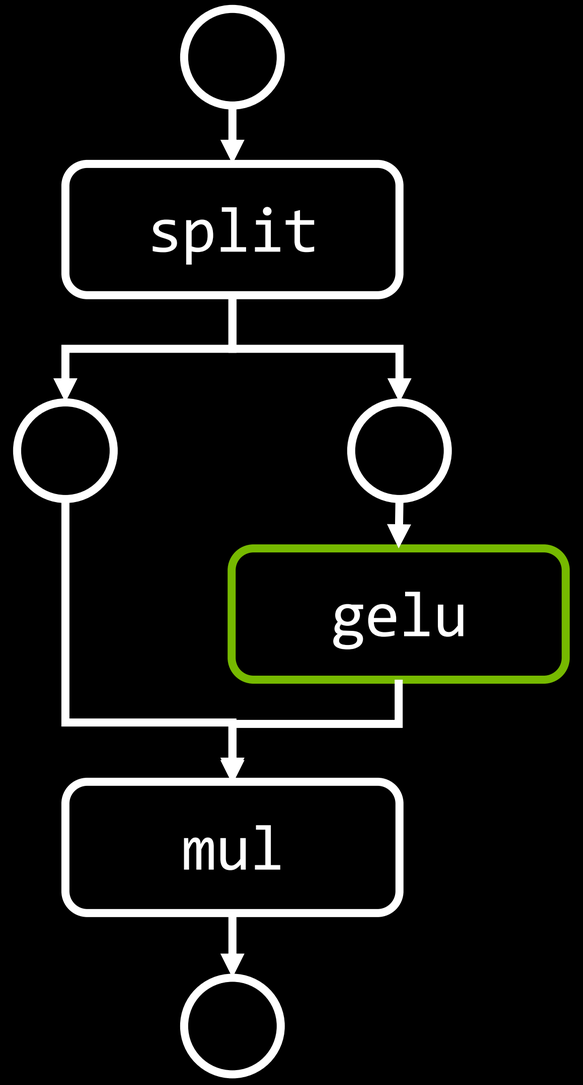


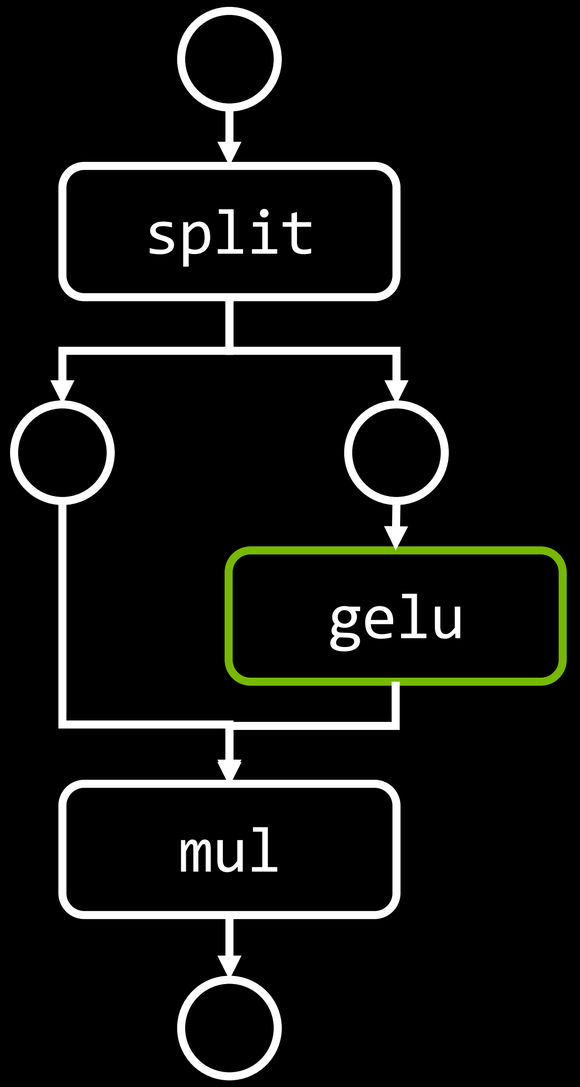


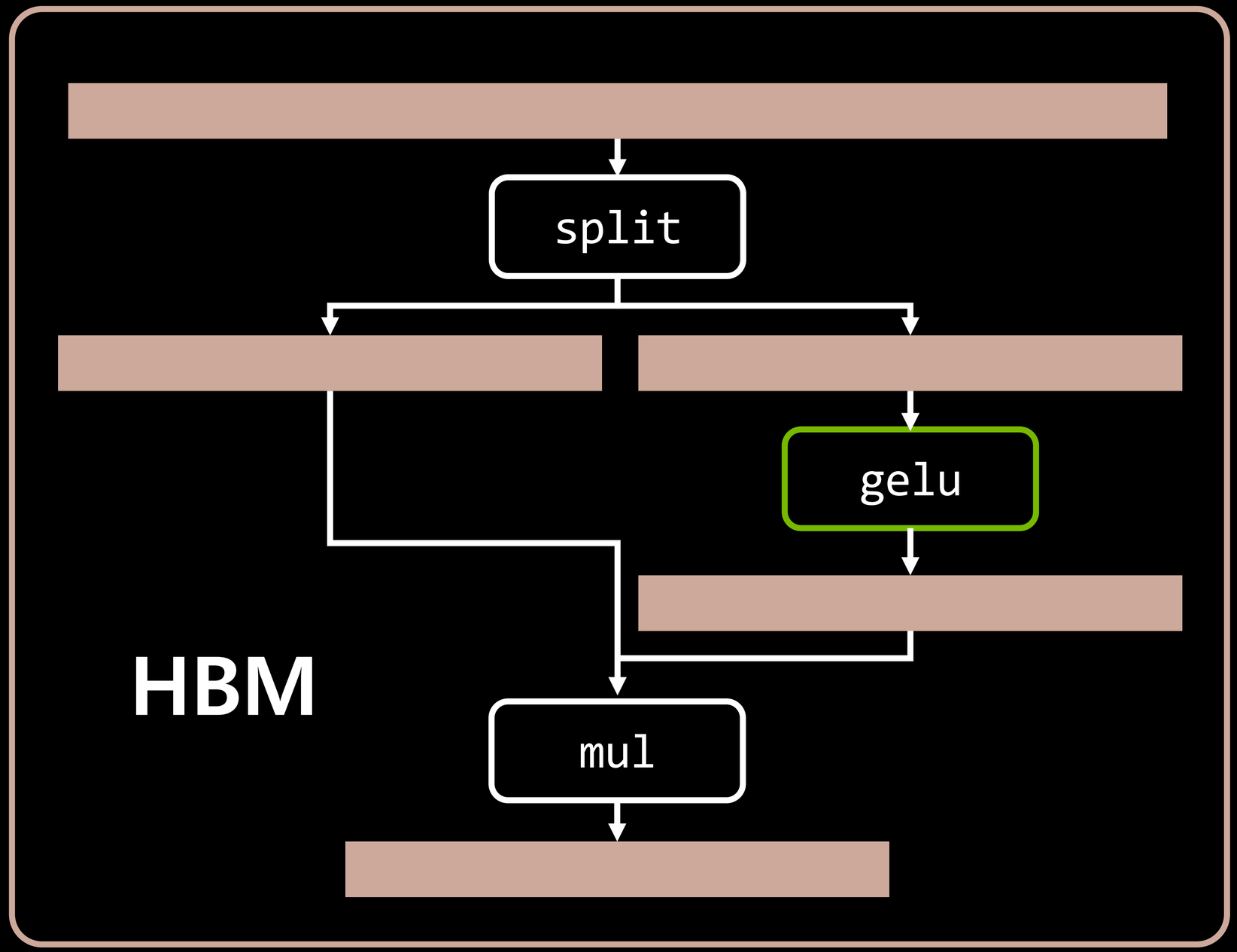
Unary operation (단항 연산자)

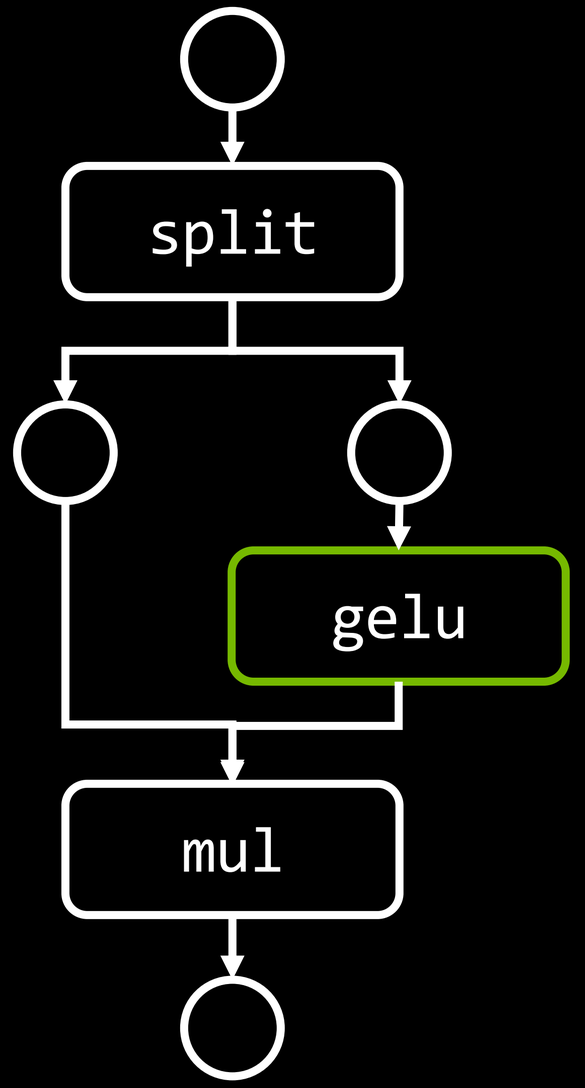
$$GELU(x) \cong 0.5x(1 + \tanh\left[\sqrt{\frac{2}{\pi}}(x + 0.4x^3)\right]$$

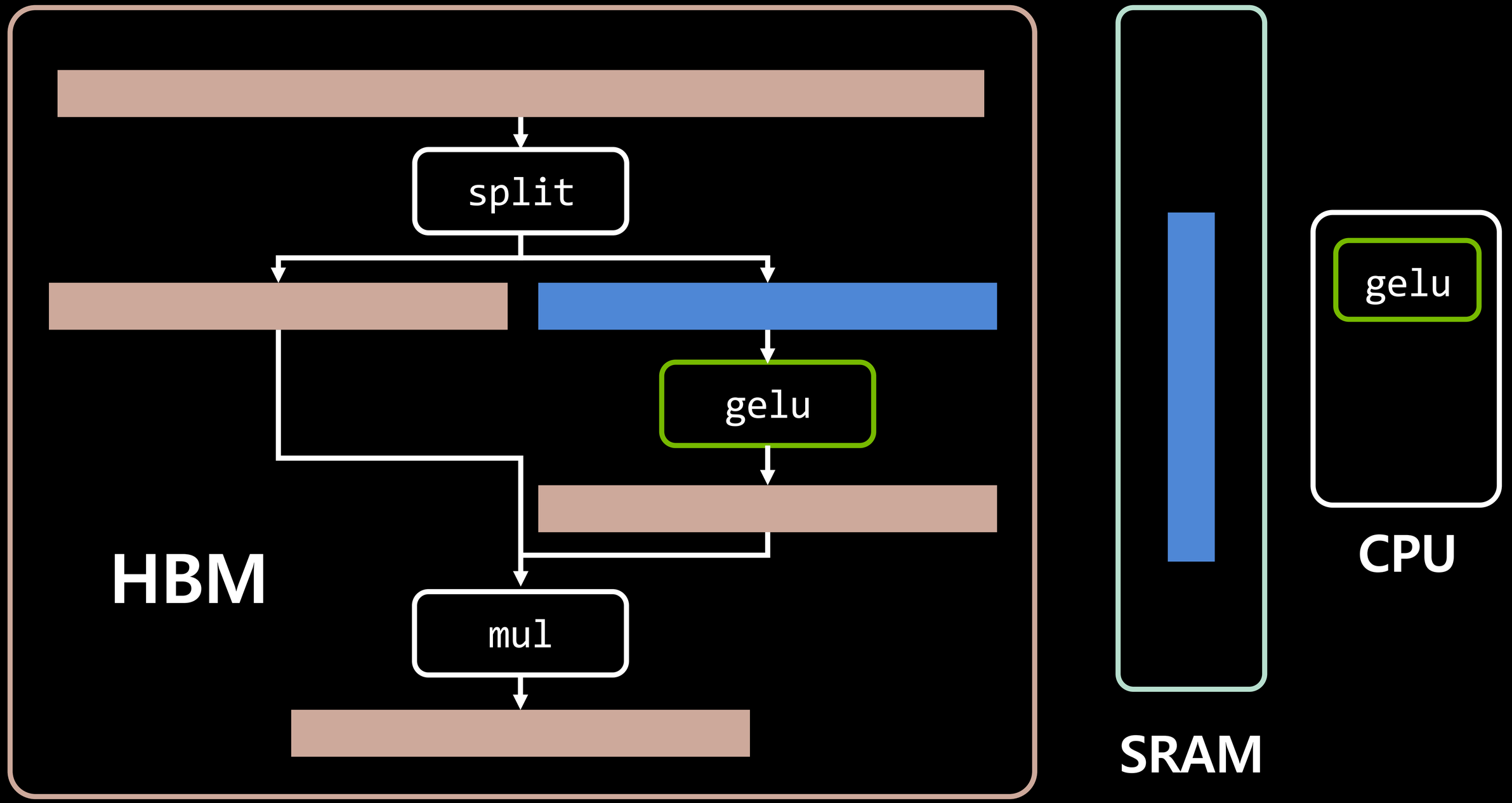


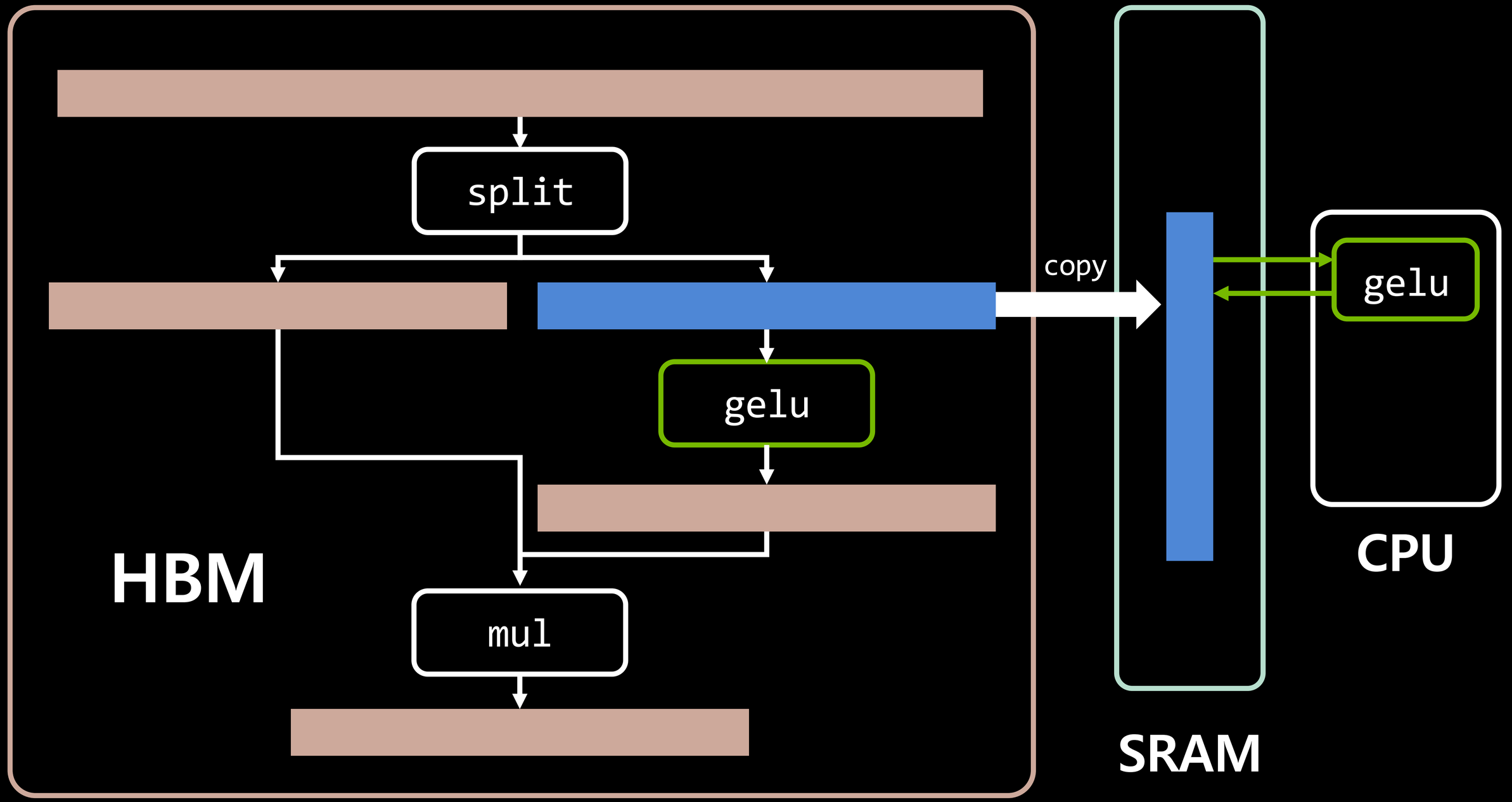


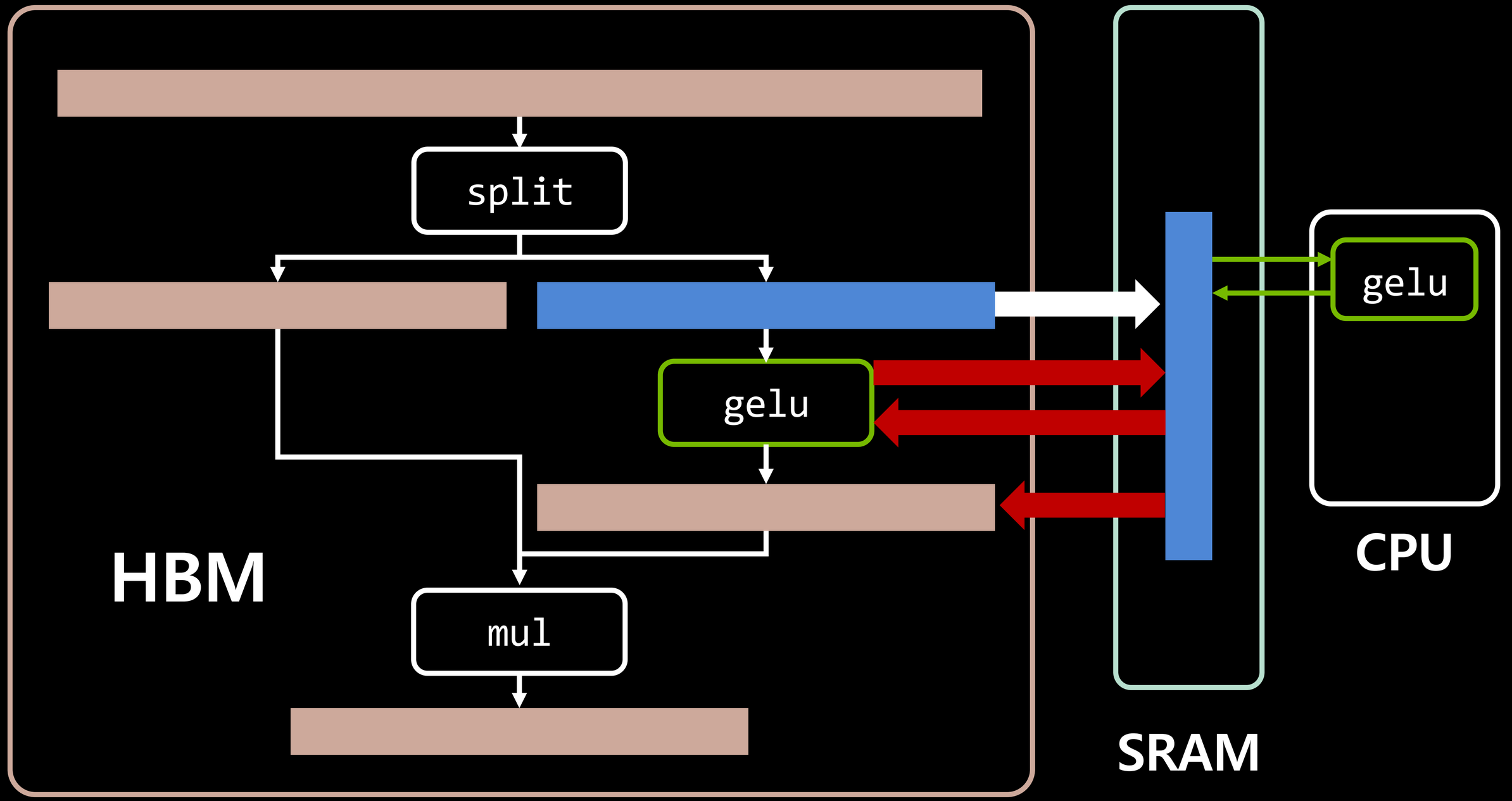

split
gelu
mul

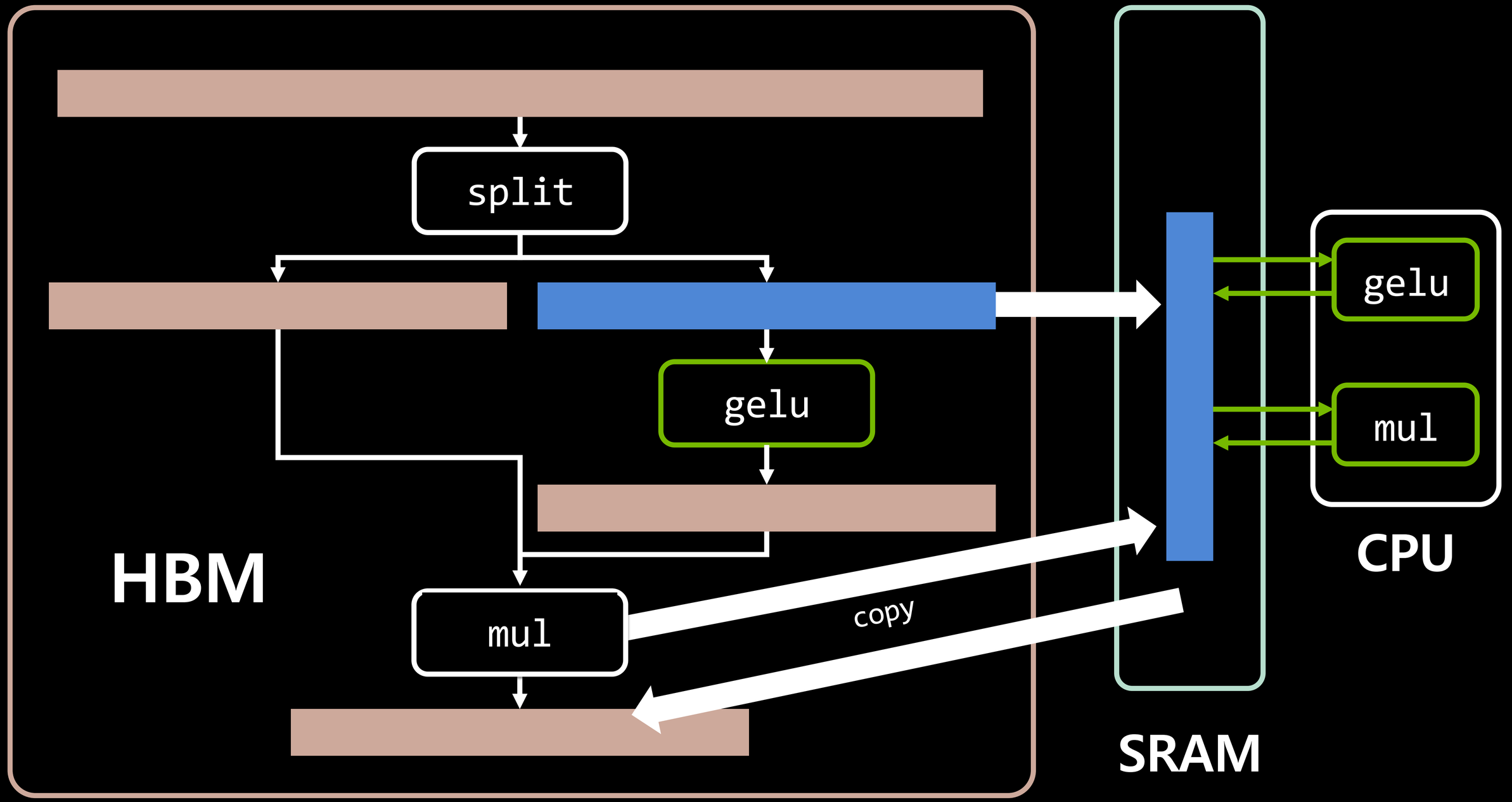

split
gelu
HBM
mul


split
gelu
mul


split
gelu
mul
HBM
SRAM
gelu
CPU


split
gelu
mul
HBM
copy
SRAM
gelu
CPU

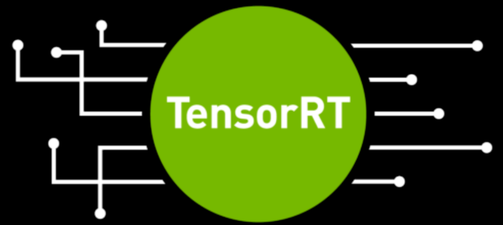

split
gelu
mul
HBM
SRAM
gelu
CPU


split
gelu
mul
copy
HBM
SRAM
gelu
mul
CPU

I/O가 7번 일어날 걸

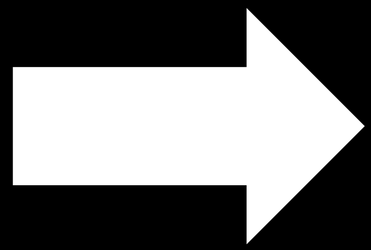
I/O가 7번 일어날 걸

1번으로 줄일 수 있다



가 최적화 하지 못하는 코드를

```python
class GEGLU(nn.Module):
  def __init__(self, dim_in, dim_out):
    super().__init__()
    self.proj = nn.Linear(dim_in, dim_out * 2)

  def forward(self, x):
    x, gate = self.proj(x).chunk(2, dim=-1)
    return x * F.gelu(gate)
```



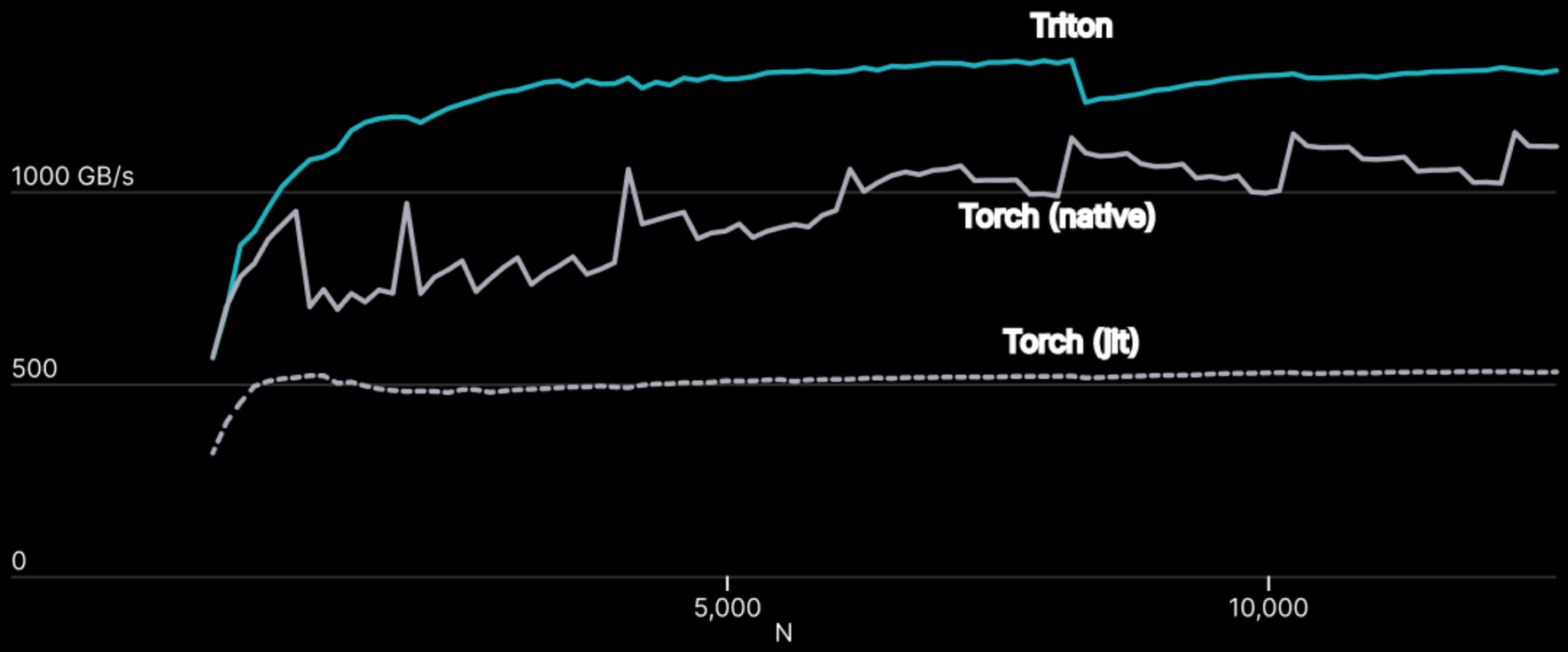
triton

OpenAI가 만든 언어 및 컴파일러

OpenAI가 만든 **언어 및 컴파일러**

I/O 최적화된 CUDA 코드를 최적화

Triton으로 짠 torch.softmax의 퍼포먼스



https://openai.com/blog/triton/

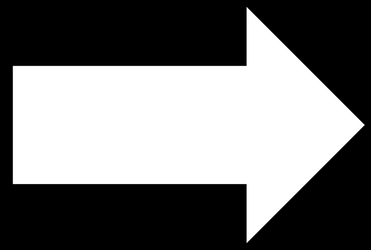
swish, gelu

Groupnorm

layernorm

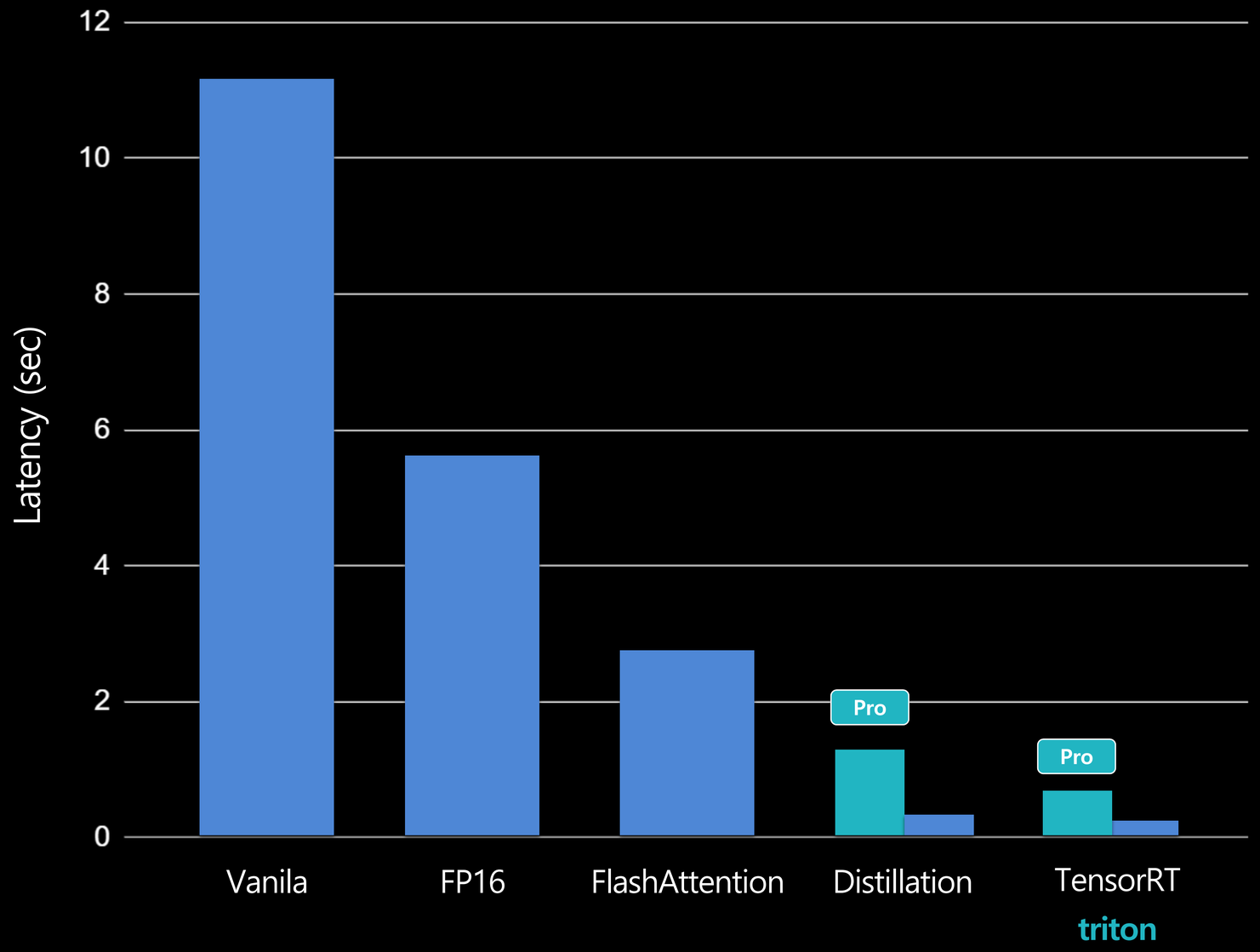
Cast

...



triton

Latency


12
10
8
6
4
2
0
Latency (sec)
Pro
Pro
Vanila
FP16
FlashAttention
Distillation
TensorRT
triton

1. Distillation by 이동익

2. TensorRT &  triton

값비싼 Diffusion model를
받드는 저비용 MLOps

1. Diffusion은 무엇이 다른가?

인프라 구조

- **Bare metal** 서버를 사용
- Kubespray로 **self-hosted** 클러스터 운영
- 2개의 서비스와 8개의 GPU 서버로 구성

Kubespray

TRITON 의 장점

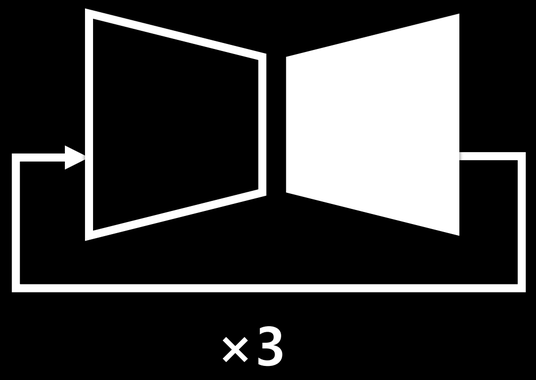
- **Ensemble Models 지원과 dlpack**
- TensorRT
- GRPC
- Dynamic Batching

TRITON INFERNCE SERVER

TRITON 의 장점

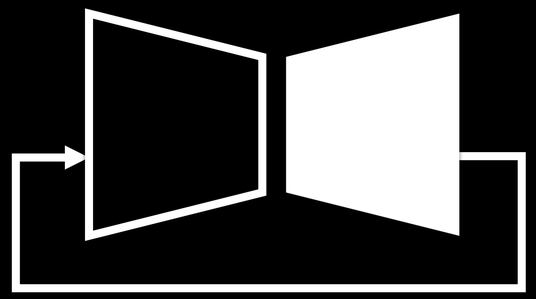
- **Ensemble Models, Scheduler 및 dlpack**
- TensorRT
- GRPC
- Dynamic Batching

TRITON INFERNCE SERVER


×3

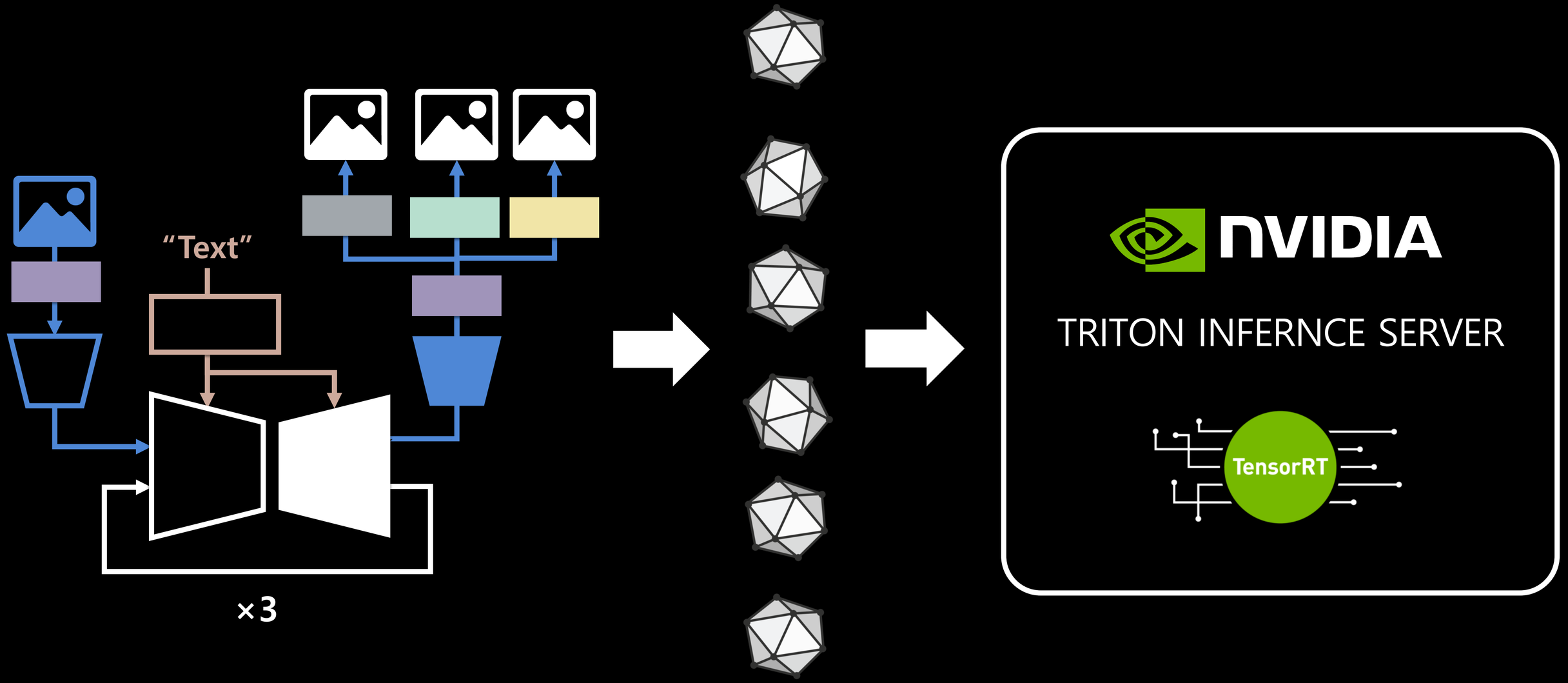
Latency

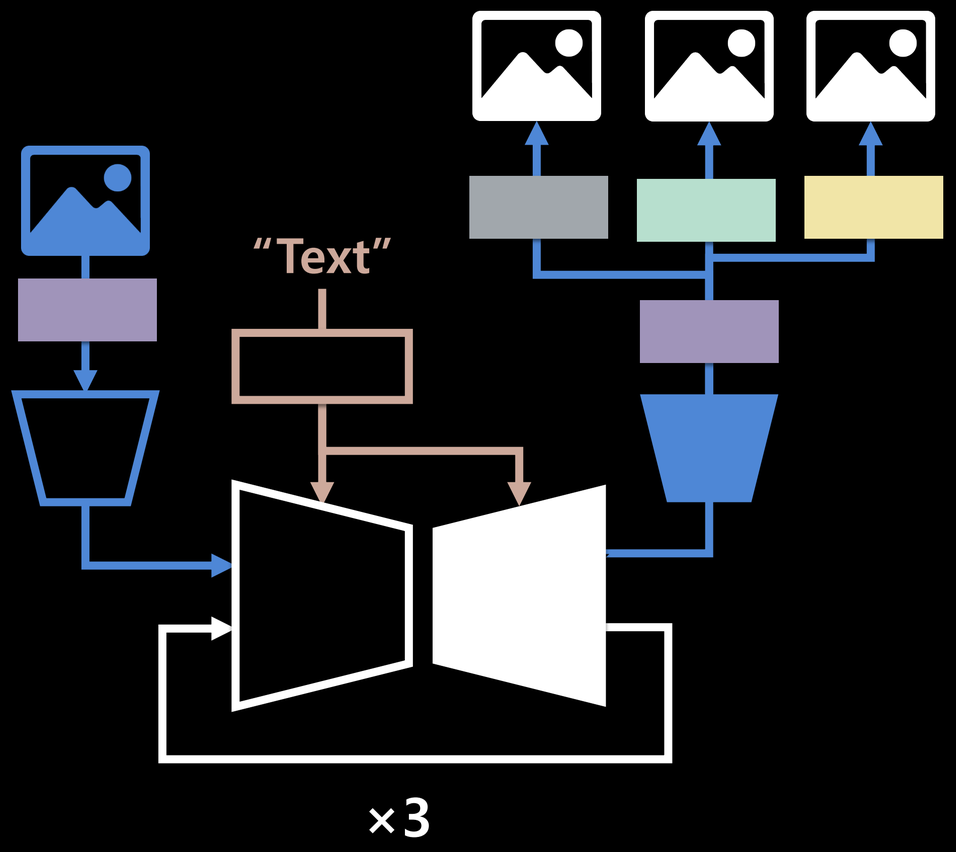
12
10
8
6
4
2
0
Latency (sec)
Pro
Pro
Vanila
FP16
FlashAttention
Distillation
TensorRT
triton

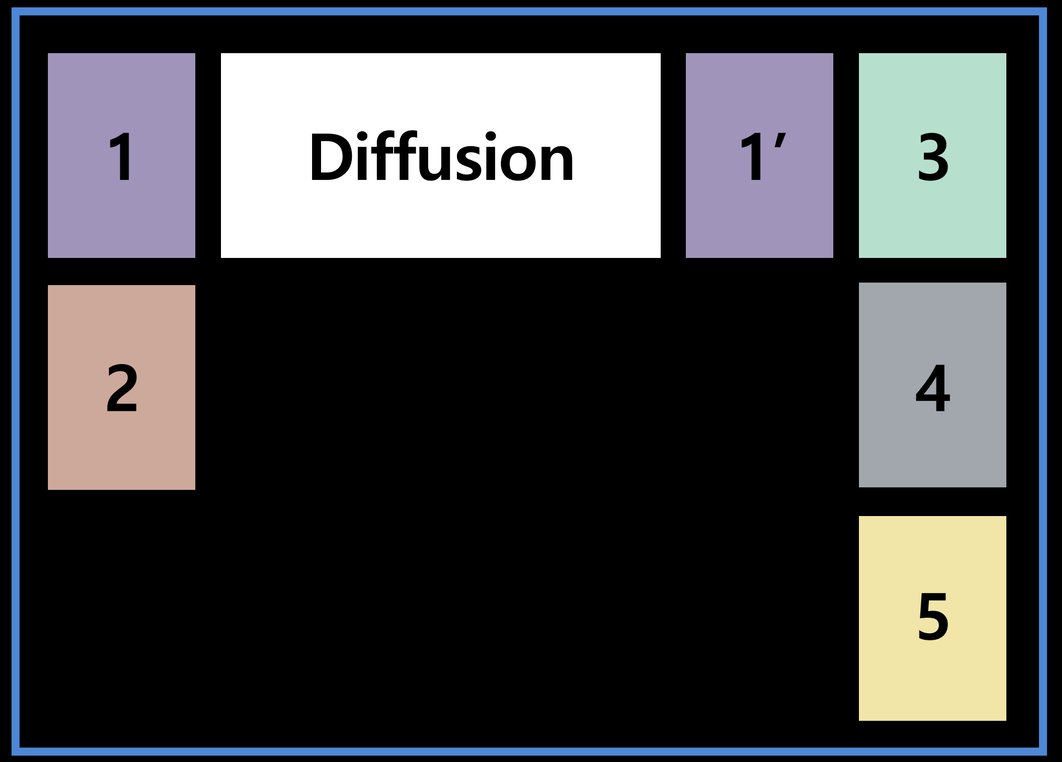


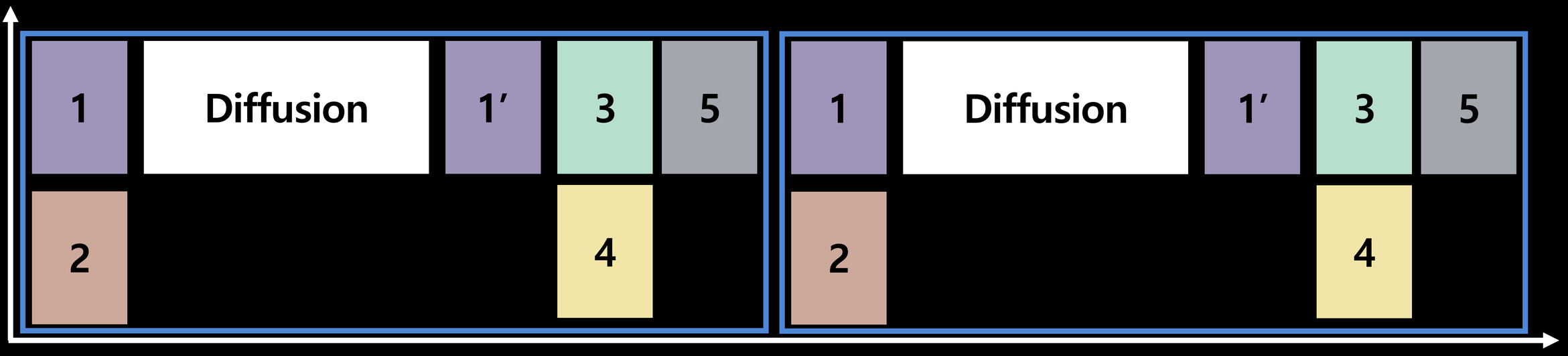
×3

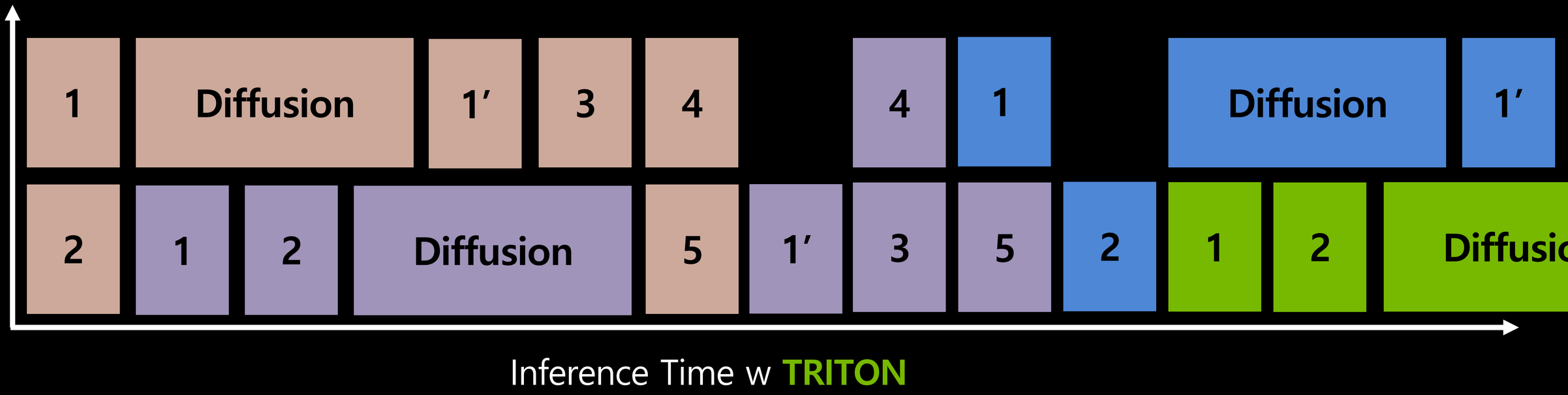
왜 이렇게까지..?


"Text"
×3
NVIDIA
TRITON INFERNCE SERVER
TensorRT


“Text”
×3


1
Diffusion
1’
3
2
4
5

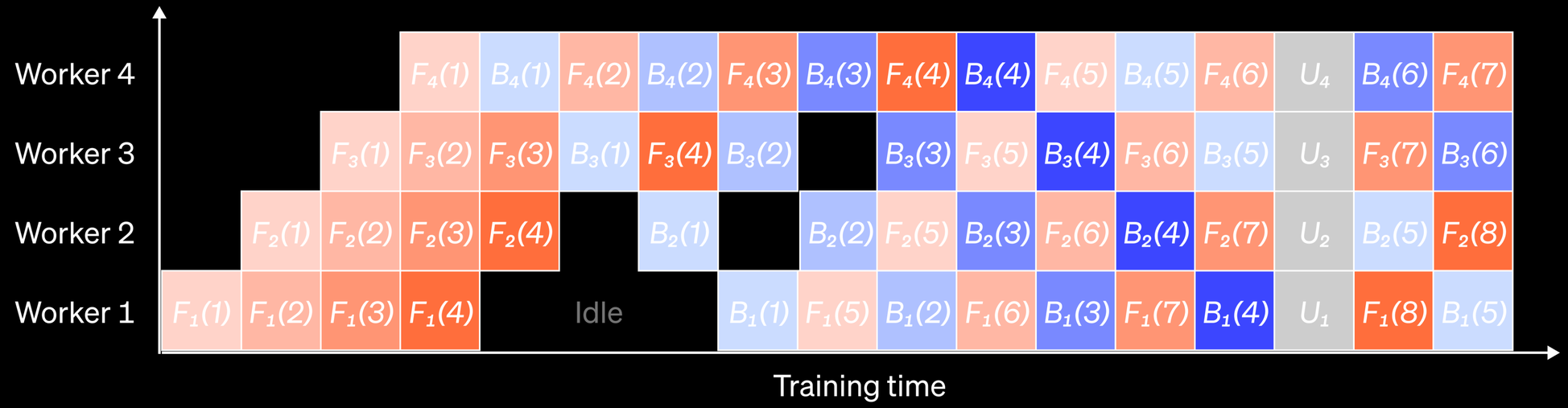

1
Diffusion
1’
3
5
2
4
1
Diffusion
1’
3
5
2
4
Inference Time


1
Diffusion
1’
3
4
4
1
Diffusion
1’
2
1
2
Diffusion
5
1’
3
5
2
1
2
Diffusio
Inference Time w TRITON

OpenAI

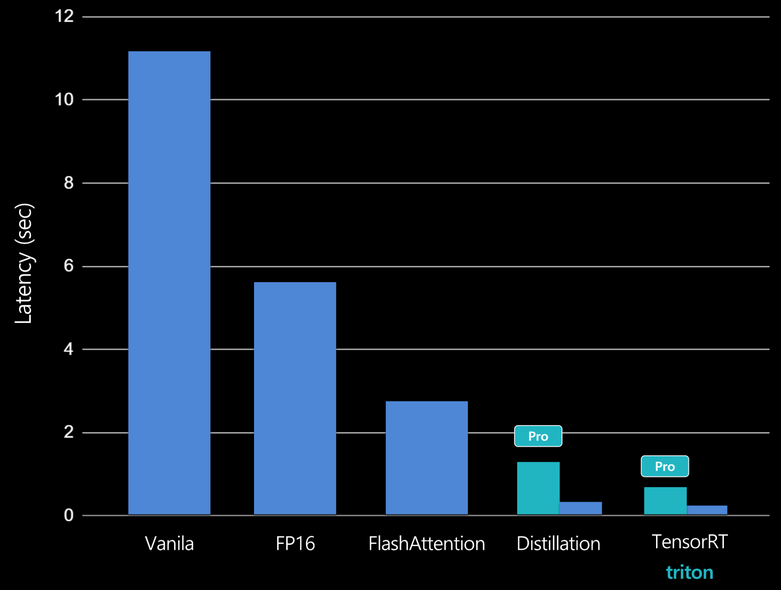
Techniques for Training Large Neural Networks

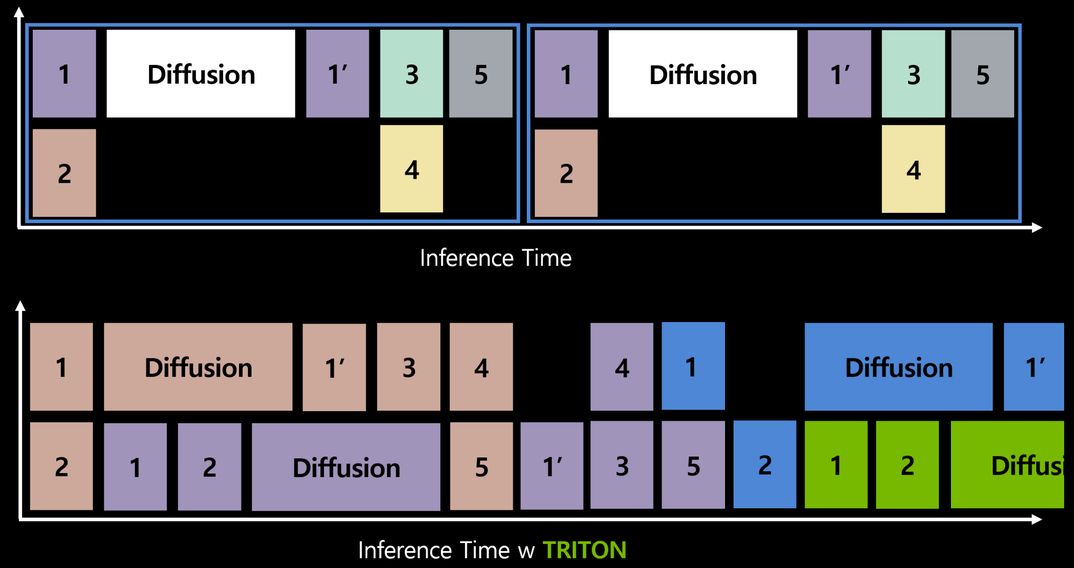
Large neural networks are at the core of many recent advances in AI, but training them is a difficult engineering and research challenge



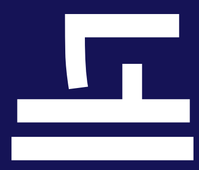
https://openai.com/blog/techniques-for-training-large-neural-networks/

Diffusion 모델과 앞 뒤로 한 두 개의 pipeline만 있다면 필요하진 않다





Latency × Scalability



값비싼 Diffusion model를 받드는 저비용 MLOps

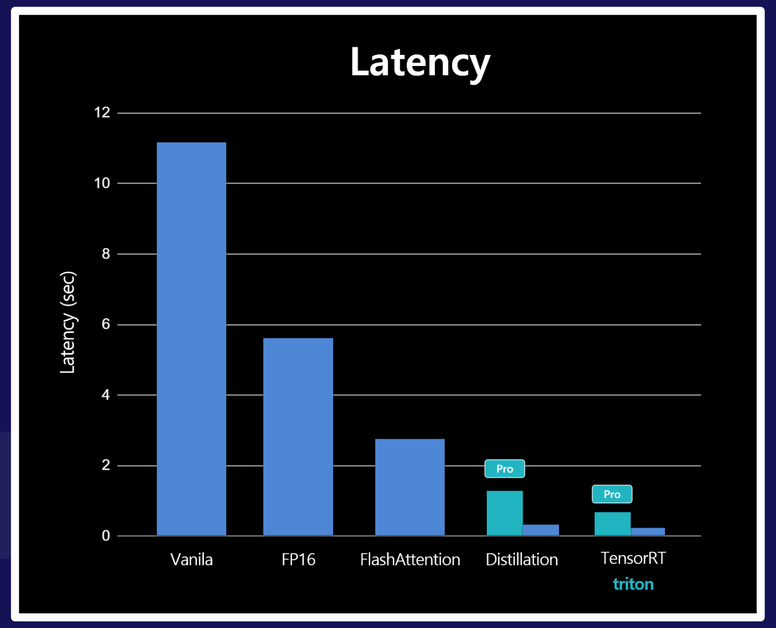
Diffusion model

Diffusion Models is

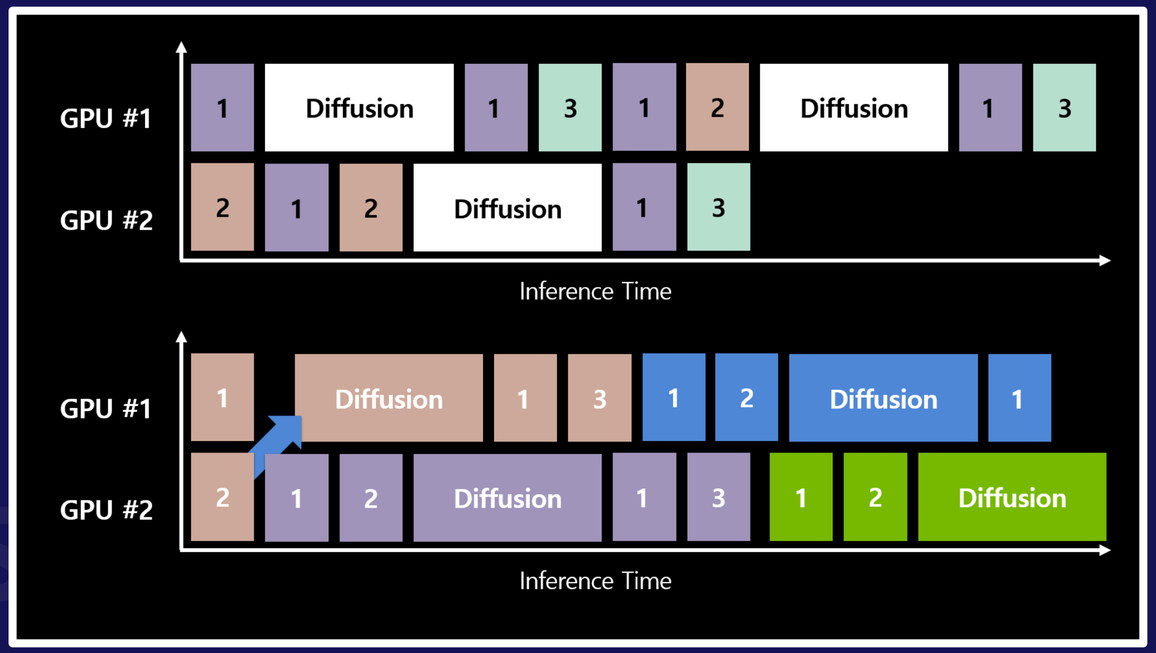
Multimodal

Image Generation model

with **diffusion process**


Latency
12
10
8
6
4
2
0
Latency (sec)
Pro
Pro
Vanila
FP16
FlashAttention
Distillation
TensorRT
triton

저비용



값비싼 Diffus

발드는 저비용 **MLOps**

마지막으로

SYMBIOTE AI

대부분의 생성 모델 스타트업

쉬운 접근성 ➡ 수많은 경쟁

AI의 대한 이해

시장과 유저

SYMBIOTE AI

글로벌 테크 회사

감사합니다