BitsFusion: 1.99 bits Weight Quantization of Diffusion Model

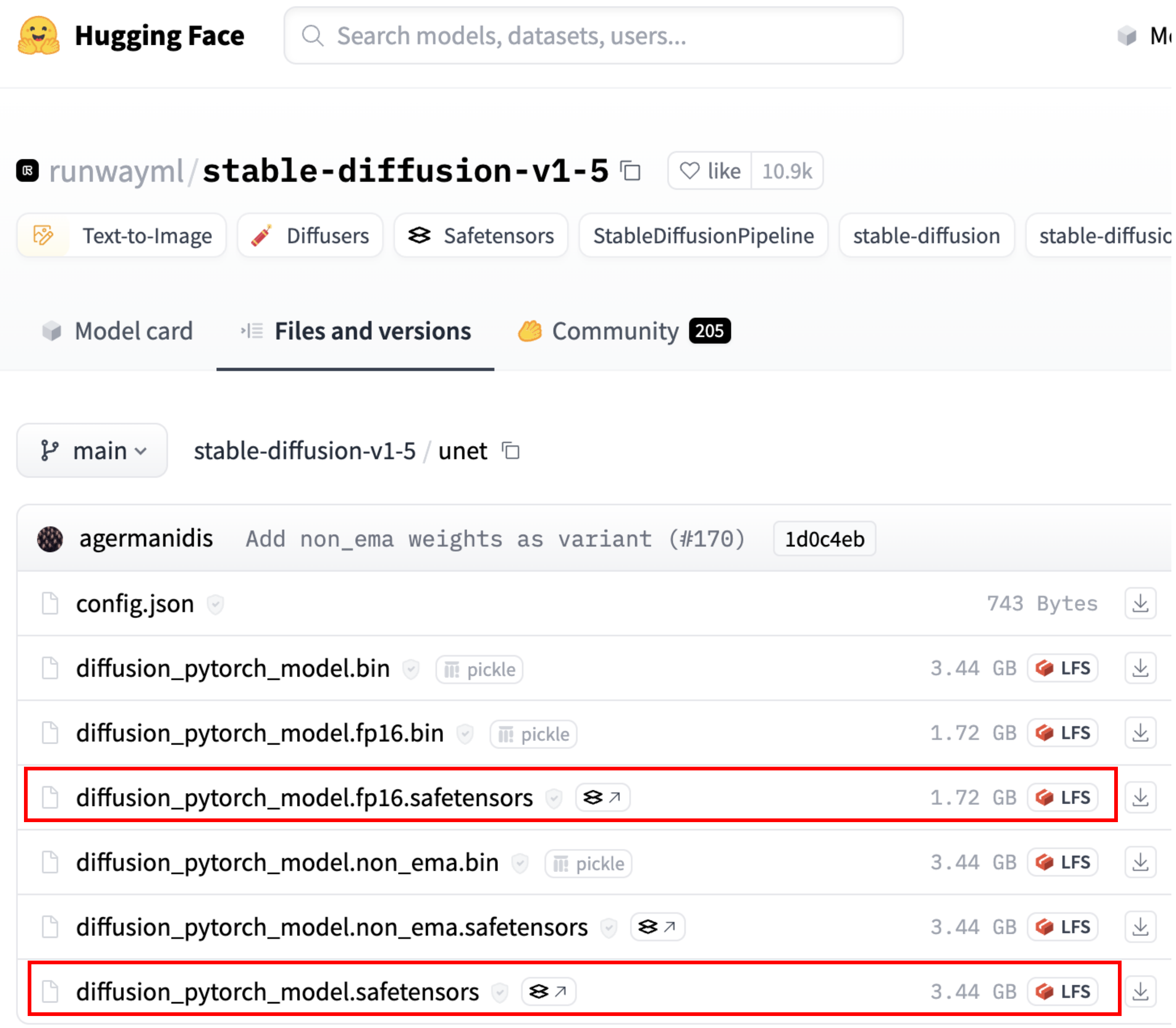
Yang Sui, Yanyu Li, Anil Kag, Yerlan Idelbayev, Junli Cao, Ju Hu,
Dhritiman Sagar, Bo Yuan, Sergey Tulyakov, Jian Ren

Snap Inc. Rutgers University

Text-to-Image Diffusion Model

Stable Diffusion

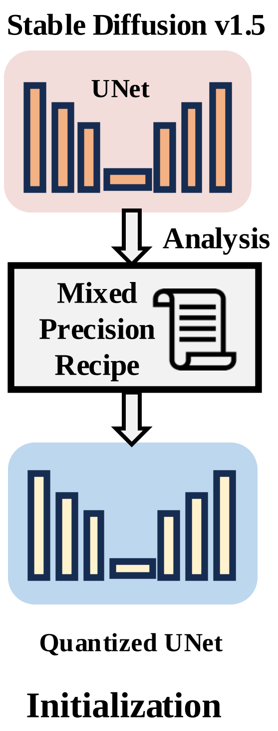
Challenge: Storage Size



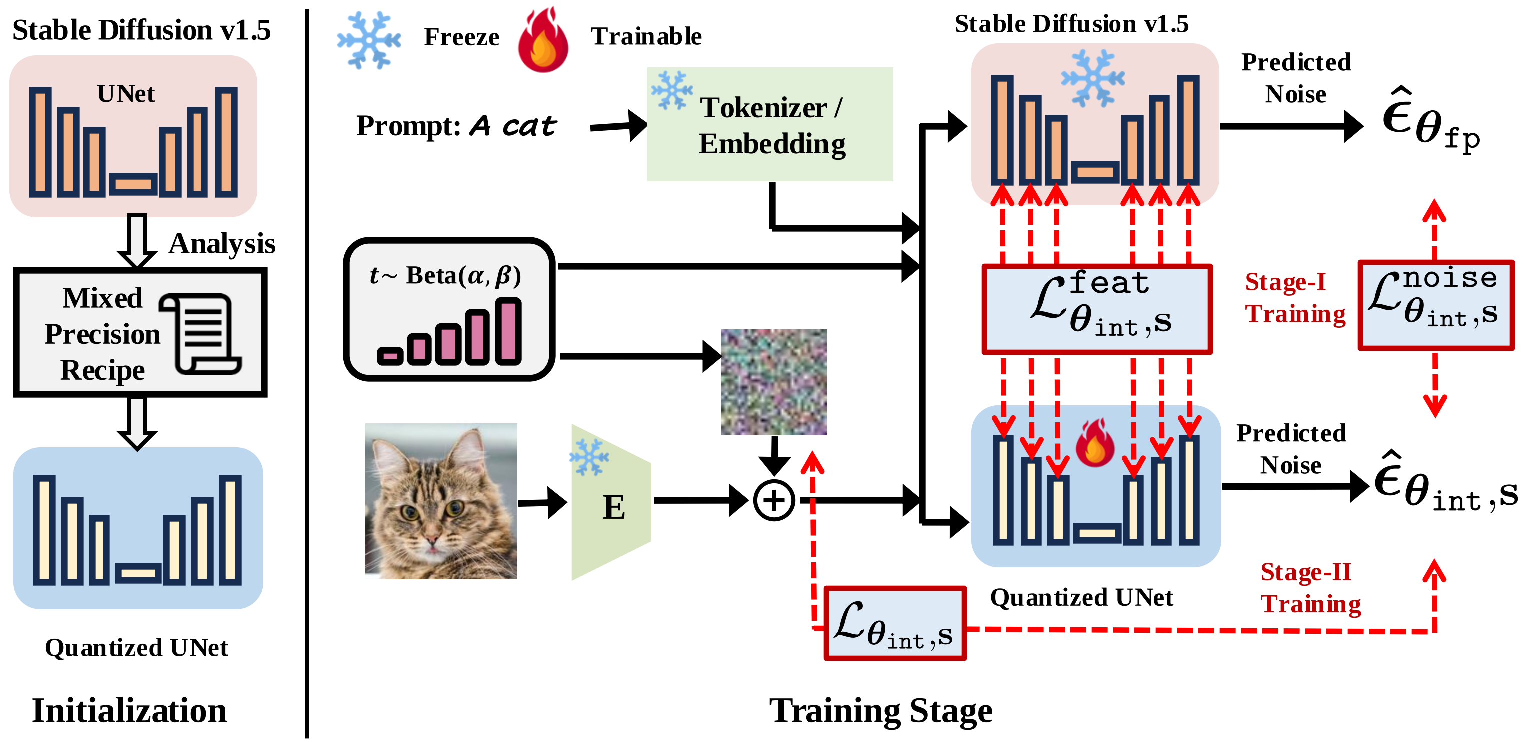
Quantization

Our goal: Extremely Low-bit Text-to-Image Diffusion Model (i.e., 1.99 bits UNet)

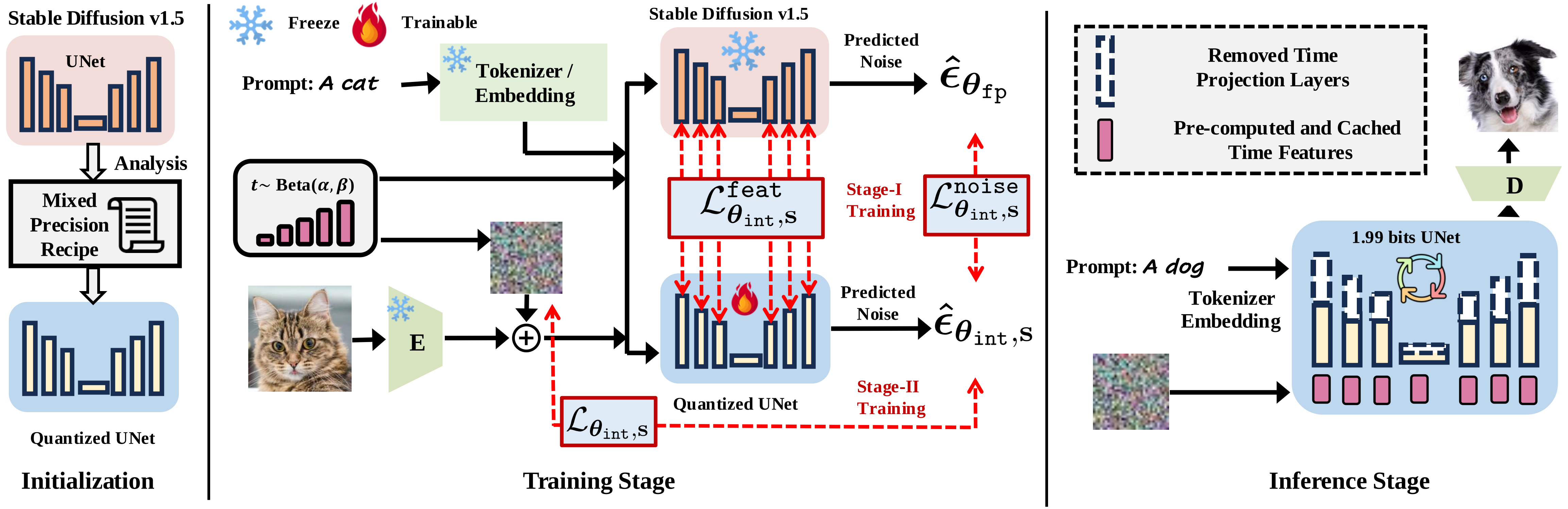
Overview of BitsFusion Pipeline



Overview of BitsFusion Pipeline



Overview of BitsFusion Pipeline



Outline

- Mixed Precision Strategy
  - Per-Layer Quantization Error Analysis
  - Deciding the Mixed Precision

- Training Extreme Low-bit Diffusion Model
  - Initialization Schemes
  - Two-stage Training Pipeline

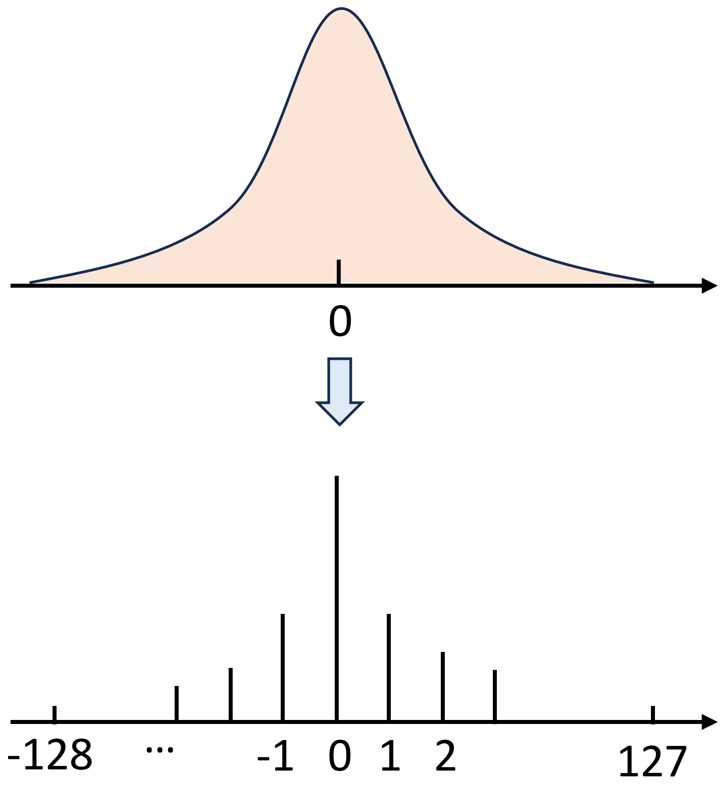
- Results

Mixed Precision Strategy

Quantization

32 bits
Floating-point Weights

8 bits
Integer Weights



Saving Storage Size

Per-Layer Quantization Error

Mixed Precision Precision

Measure the impact when quantizing each layer:

- Quantize each single layer to 1,2,3 bits

Per-Layer Quantization Error

Mixed Precision Precision

Measure the impact when quantizing each layer:

- Quantize each single layer to 1,2,3 bits
- Generate images from quantized models

Per-Layer Quantization Error

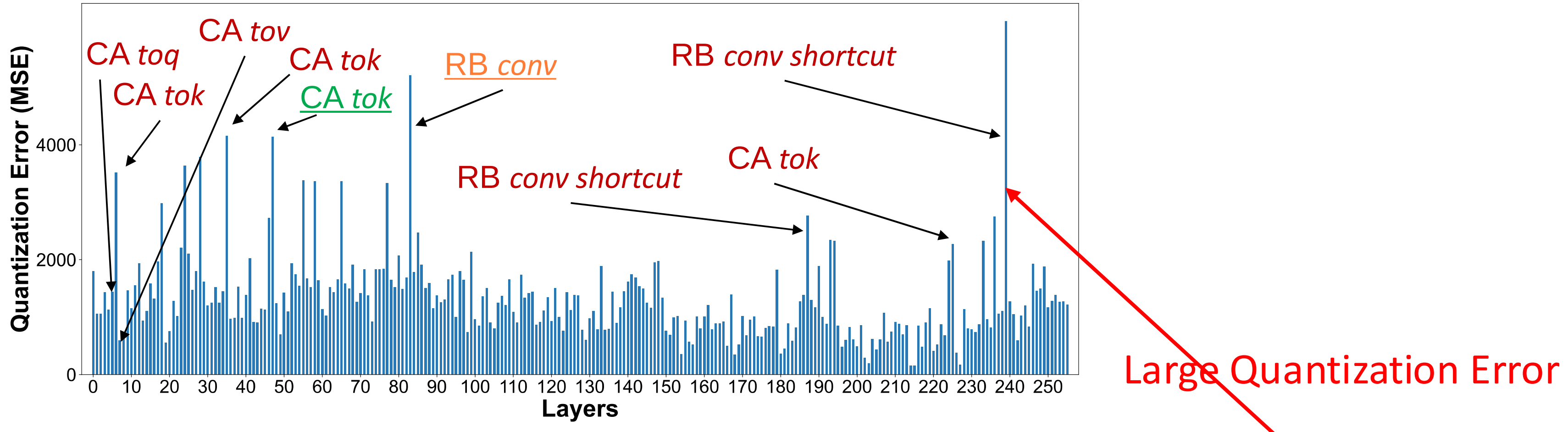
Mixed Precision Precision

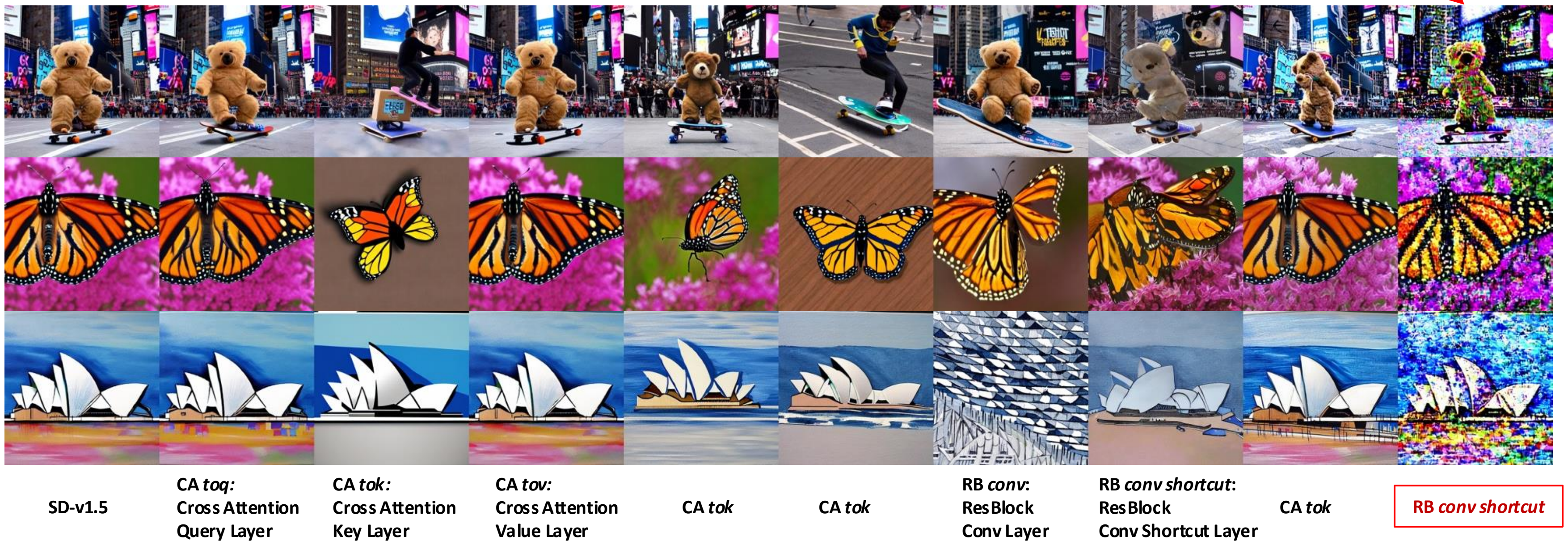
Measure the impact when quantizing each layer:

- Quantize each single layer to 1,2,3 bits
- Generate images from quantized models
- Calculate the metrics compared to full-precision model: MSE, CLIP Score, PSNR, LPIPS

Per-Layer Quantization Error

Mixed Precision Precision





Per-Layer Quantization Error

Which metrics should we use?

Pearson correlation (absolute value) between different metrics

	MSE vs. PSNR	MSE vs. LPIPS	MSE vs. CLIP Score
1 bit	0.870	0.984	0.733
2 bit	0.882	0.989	0.473
3 bit	0.869	0.991	0.535

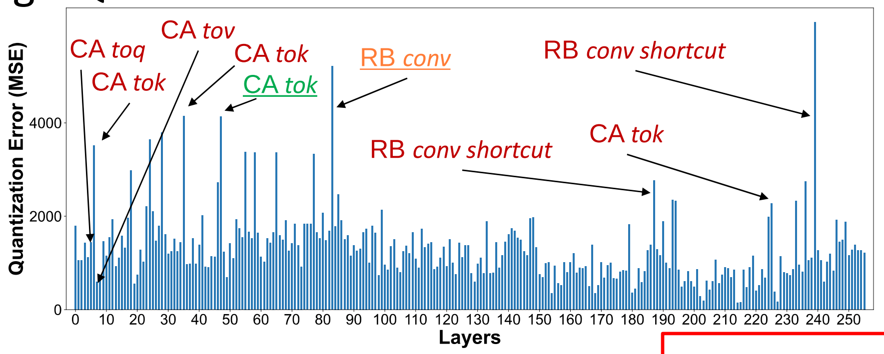
Observation 1: MSE, PSNR, and LPIPS show strong correlation and they correlate well with the visual perception of image quality.

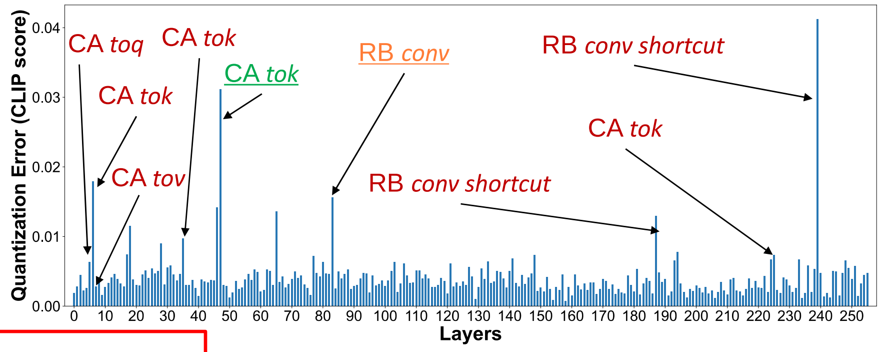
Conclusion: We adopt MSE as our main quantitative metric to represent the PSNR and LPIPS.

Per-Layer Quantization Error

Which metrics should we use?

Average Quantization Error:





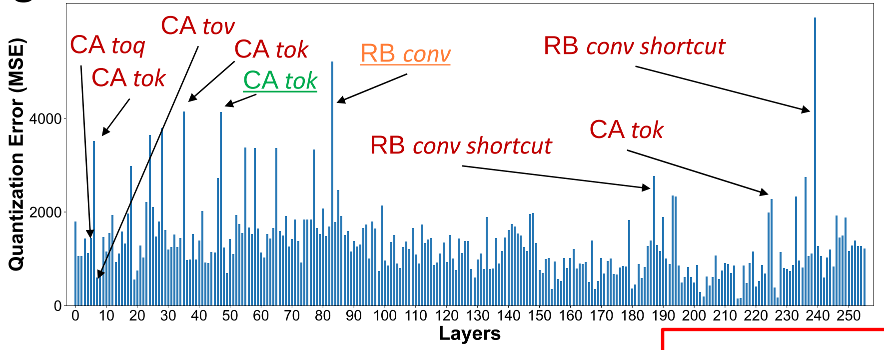
MSE: CA *tok* < RB *conv*
CLIP Score drop: CA *tok* > RB *conv*

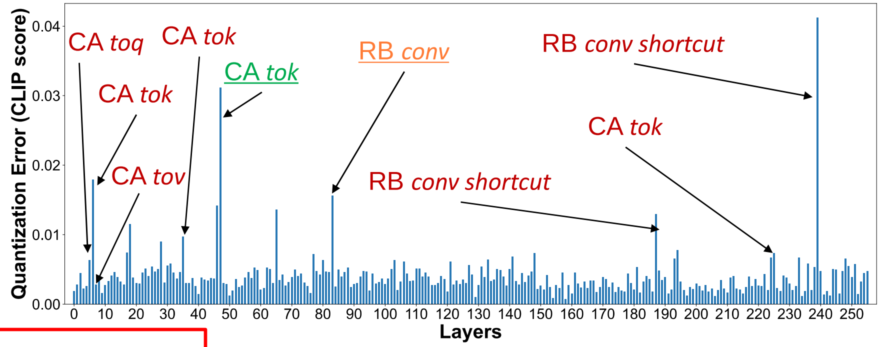
Observation 2: Although some layers show smaller MSE, they may experience larger semantic degradation, as reflected in larger CLIP score changes.

Per-Layer Quantization Error

Which metrics should we use?

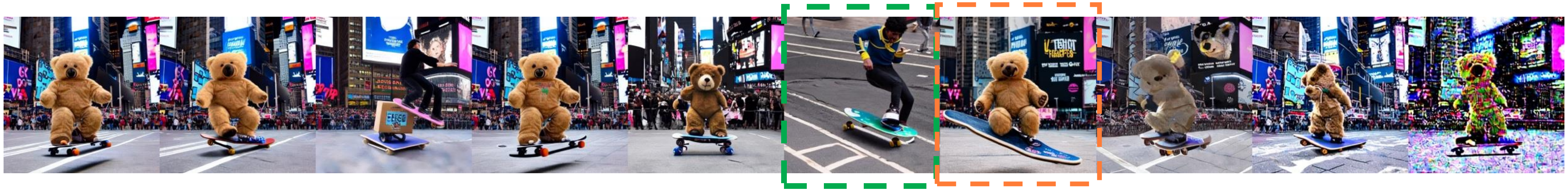
Average Quantization Error:





MSE: *CA tok* < *RB conv*
CLIP Score drop: *CA tok* > *RB conv*

Observation 2: Although some layers show smaller MSE, they may experience larger semantic degradation, as reflected in larger CLIP score changes.

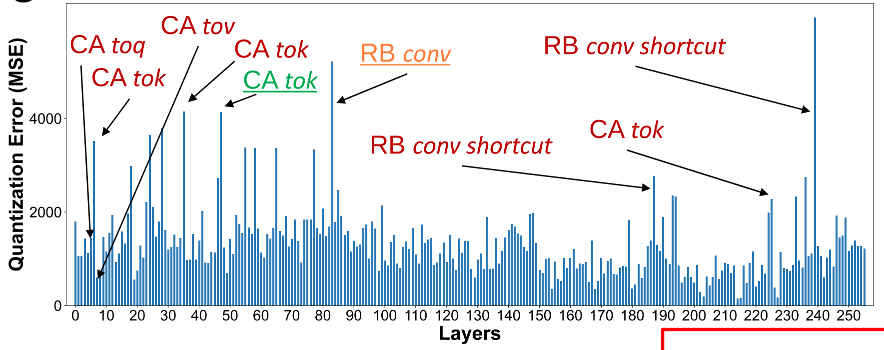


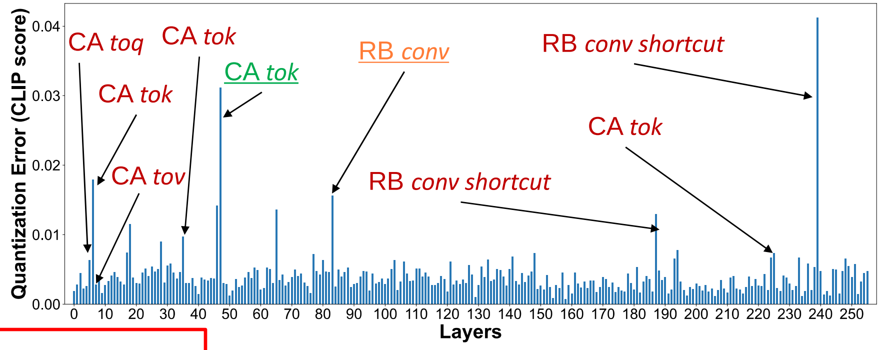
A teddy bear on a skateboard in Times Square, doing tricks on a cardboard box ramp

Per-Layer Quantization Error

Which metrics should we use?

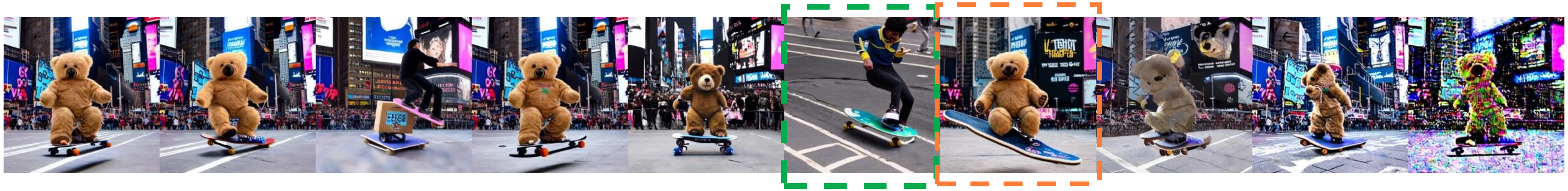
Average Quantization Error:





MSE: CA *tok* < RB *conv*
CLIP Score drop: CA *tok* > RB *conv*

Observation 2: Although some layers show smaller MSE, they may experience larger semantic degradation, as reflected in larger CLIP score changes.



A teddy bear on a skateboard in Times Square, doing tricks on a cardboard box ramp

Adopt CLIP score as our complementary quantitative metrics.

Per-Layer Quantization Error

Which metrics should we use?

Conclusion: MSE, CLIP Score

Deciding the Optimal Precision

Sensitivity score

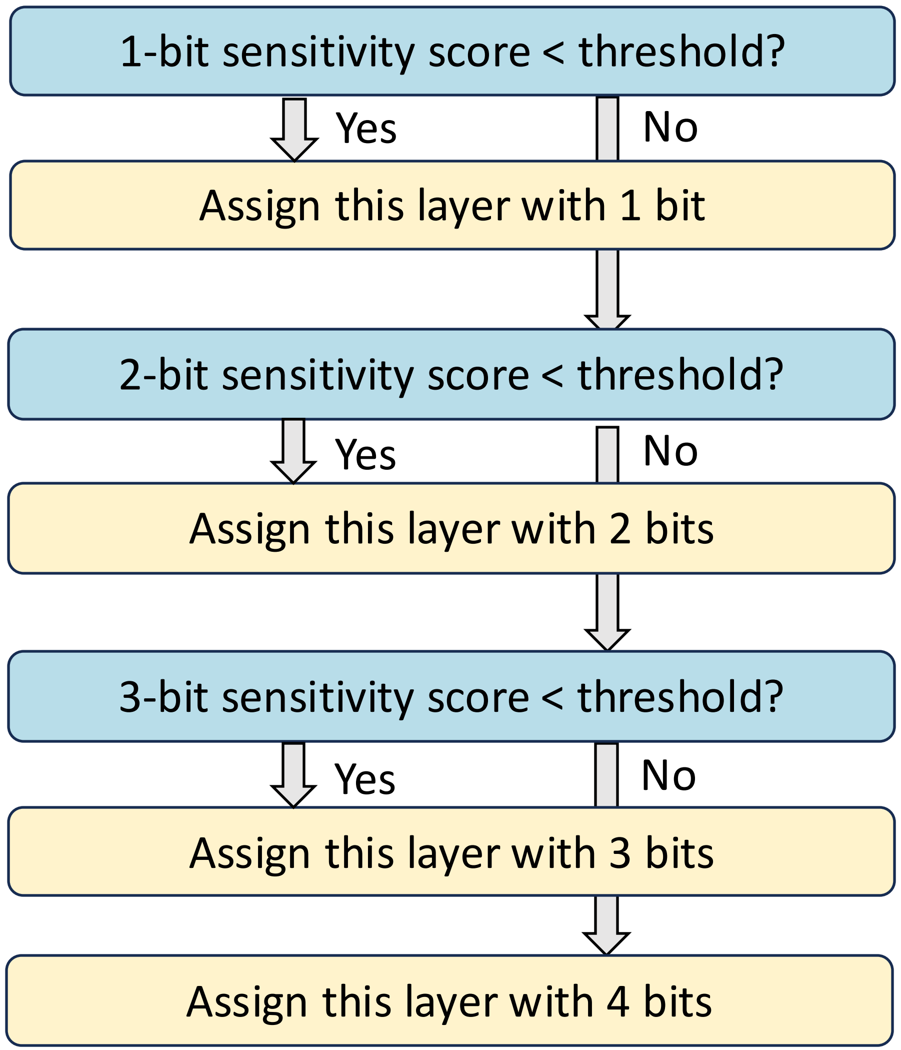
$$\mathcal{S}_{i,b} = M_{i,b} N_i^{-\eta}$$

$M_{i,b}$: MSE

N_i : Parameter size

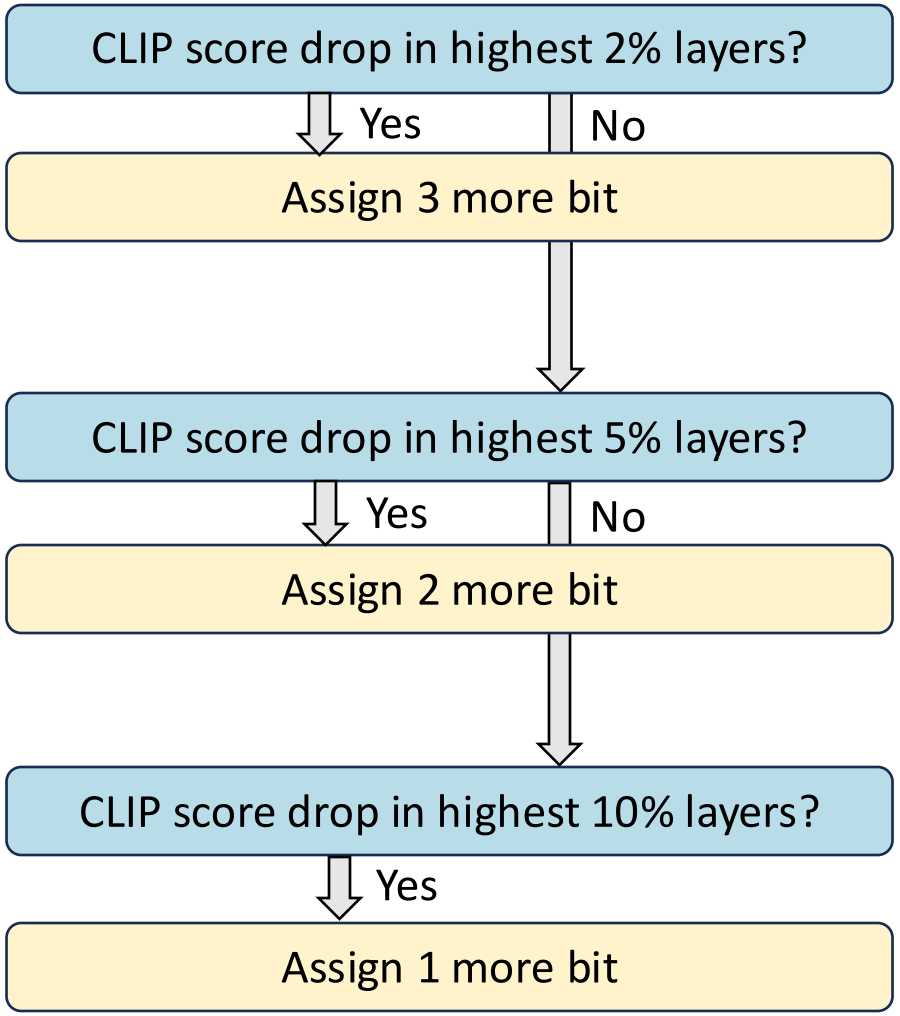
η : Parameter size factor

1. Assign bits based on MSE (For one layer):



Deciding the Optimal Precision

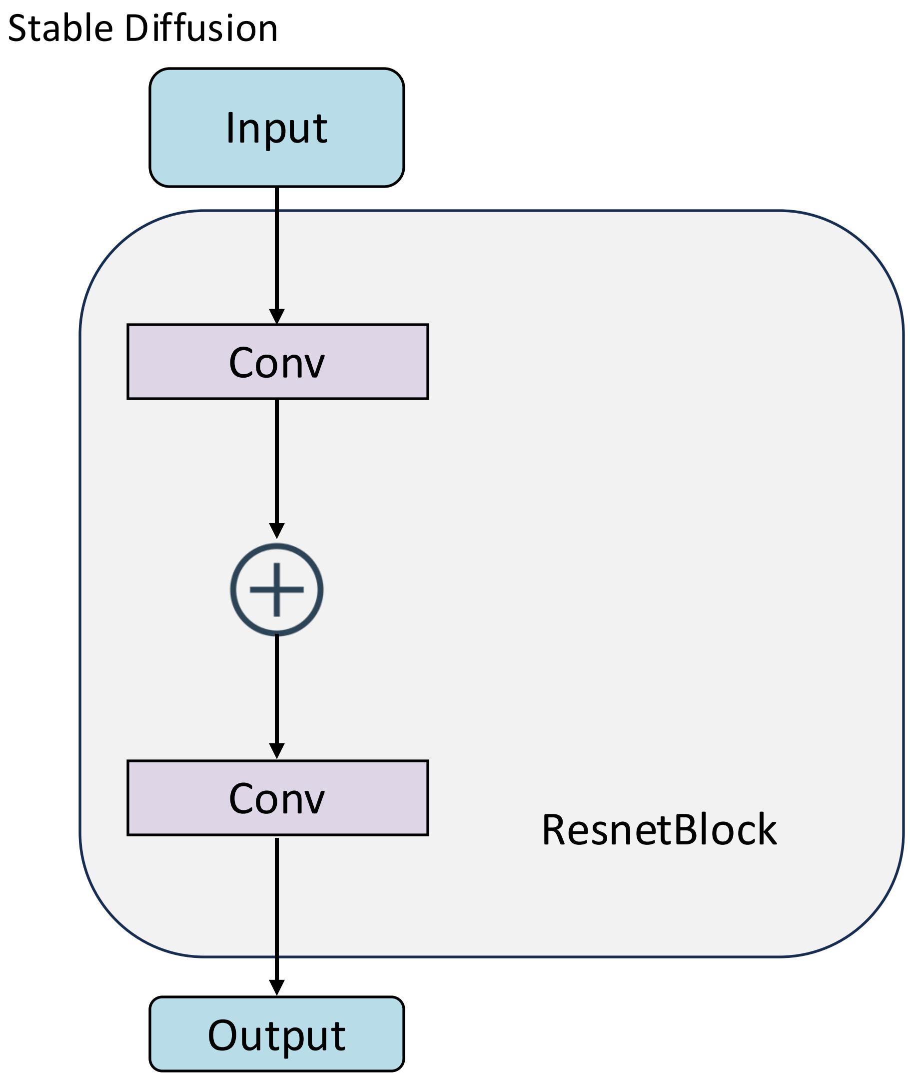
2. Adjust bits based on CLIP scores (For one layer):



Initialization

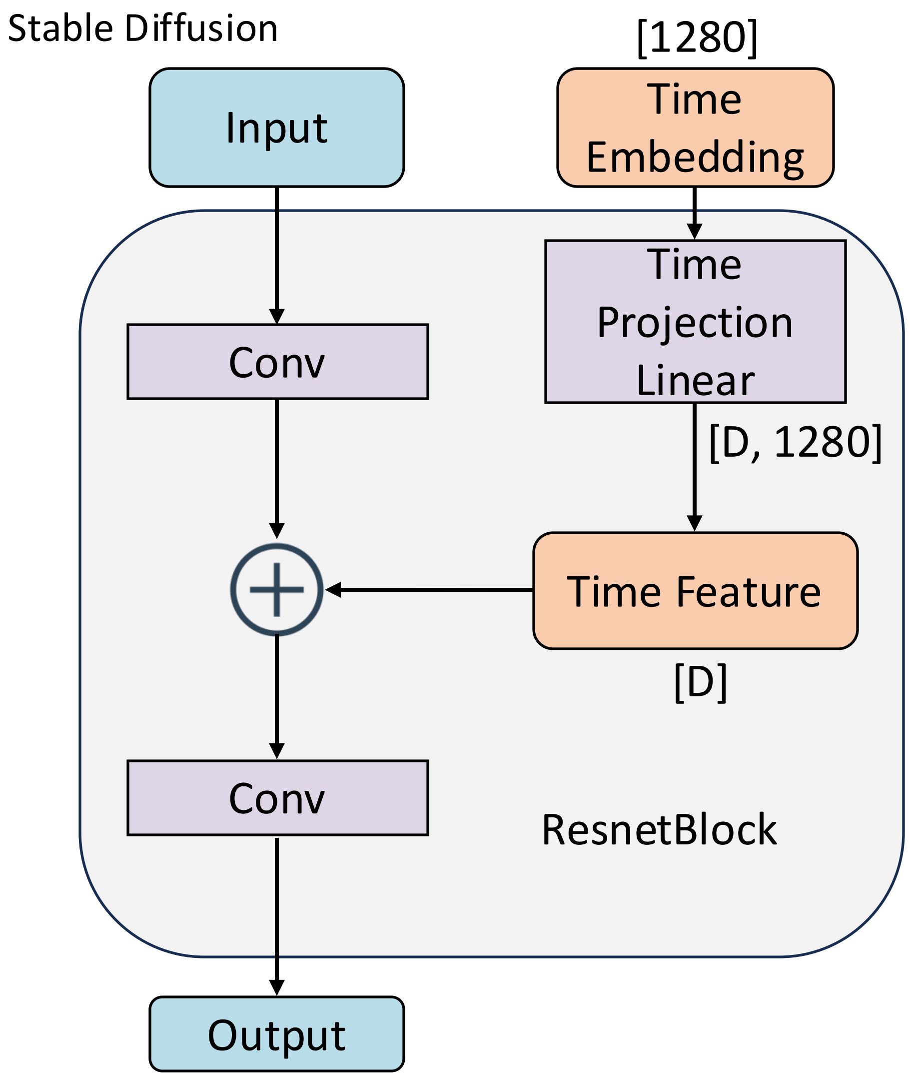
Initialization

Time Embedding Pre-computing and Caching



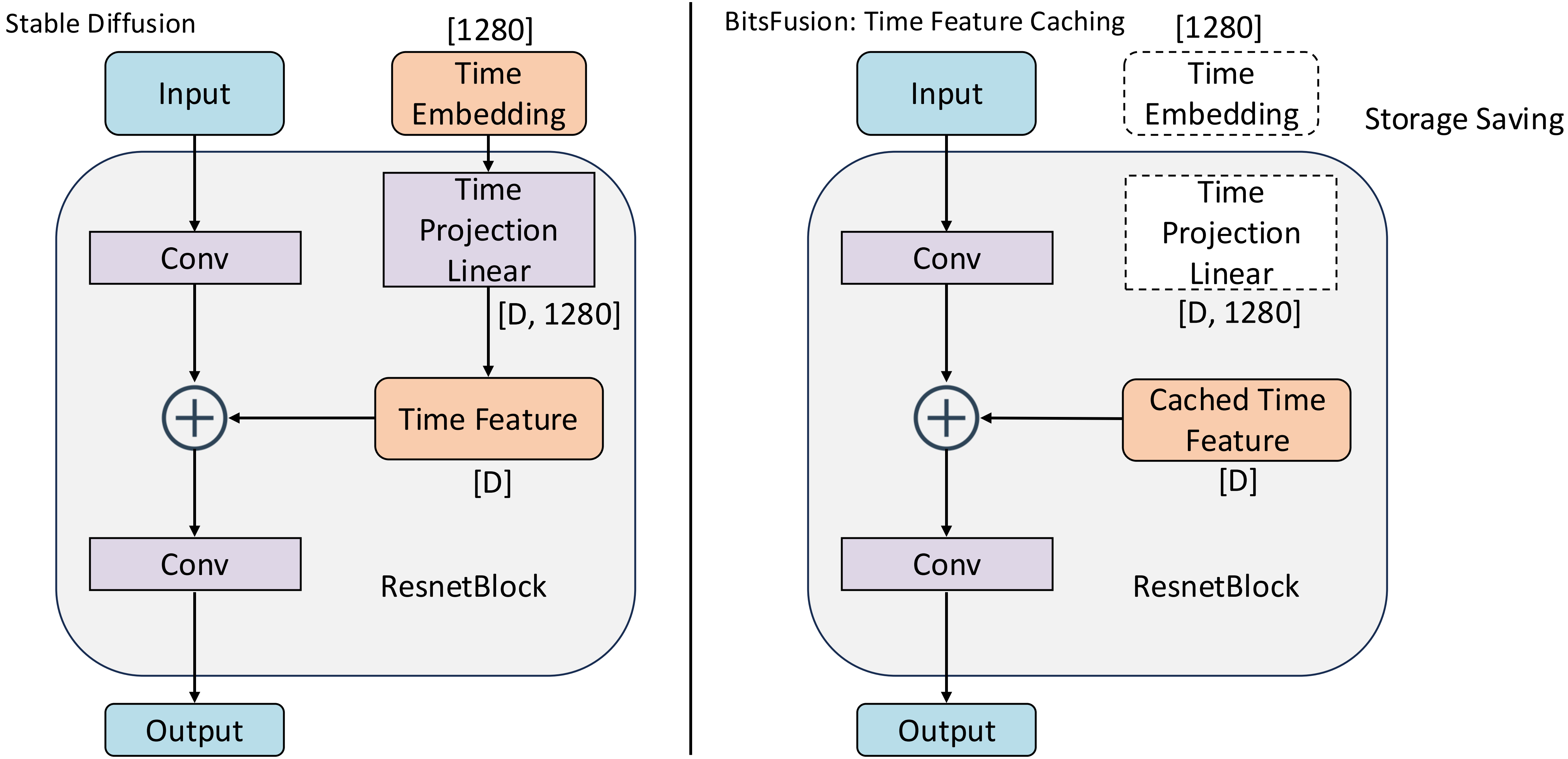
Initialization

Time Embedding Pre-computing and Caching



Initialization

Time Embedding Pre-computing and Caching

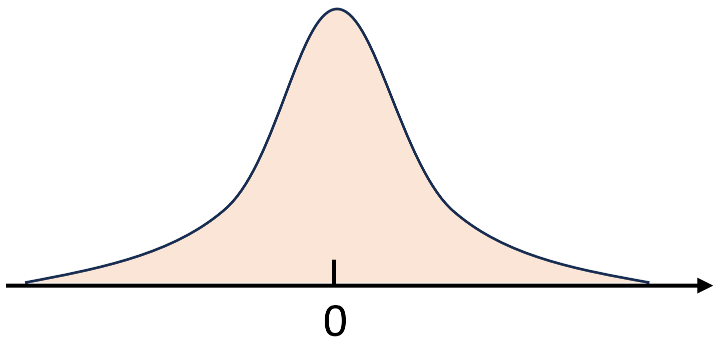


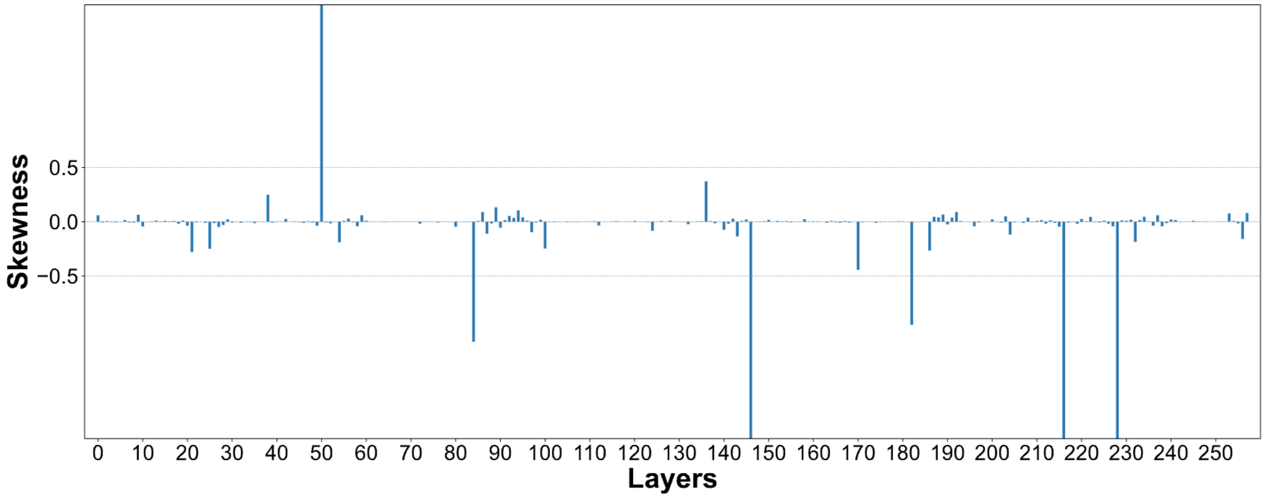
During inference stage, only caching T time features (T is the sampling steps, T <= 50 in stable diffusion).

Initialization

Adding Balance Integer

Is weight distribution symmetric in Stable Diffusion?





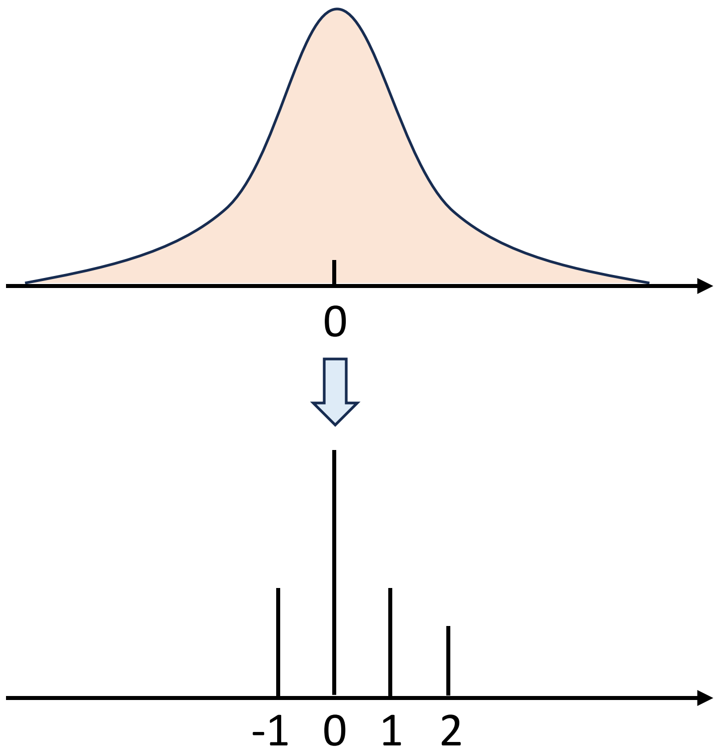
97% of layers exhibiting skewness between [-0.5, 0.5]

Weight Distribution of layers in Stable Diffusion is symmetric

Initialization

Adding Balance Integer

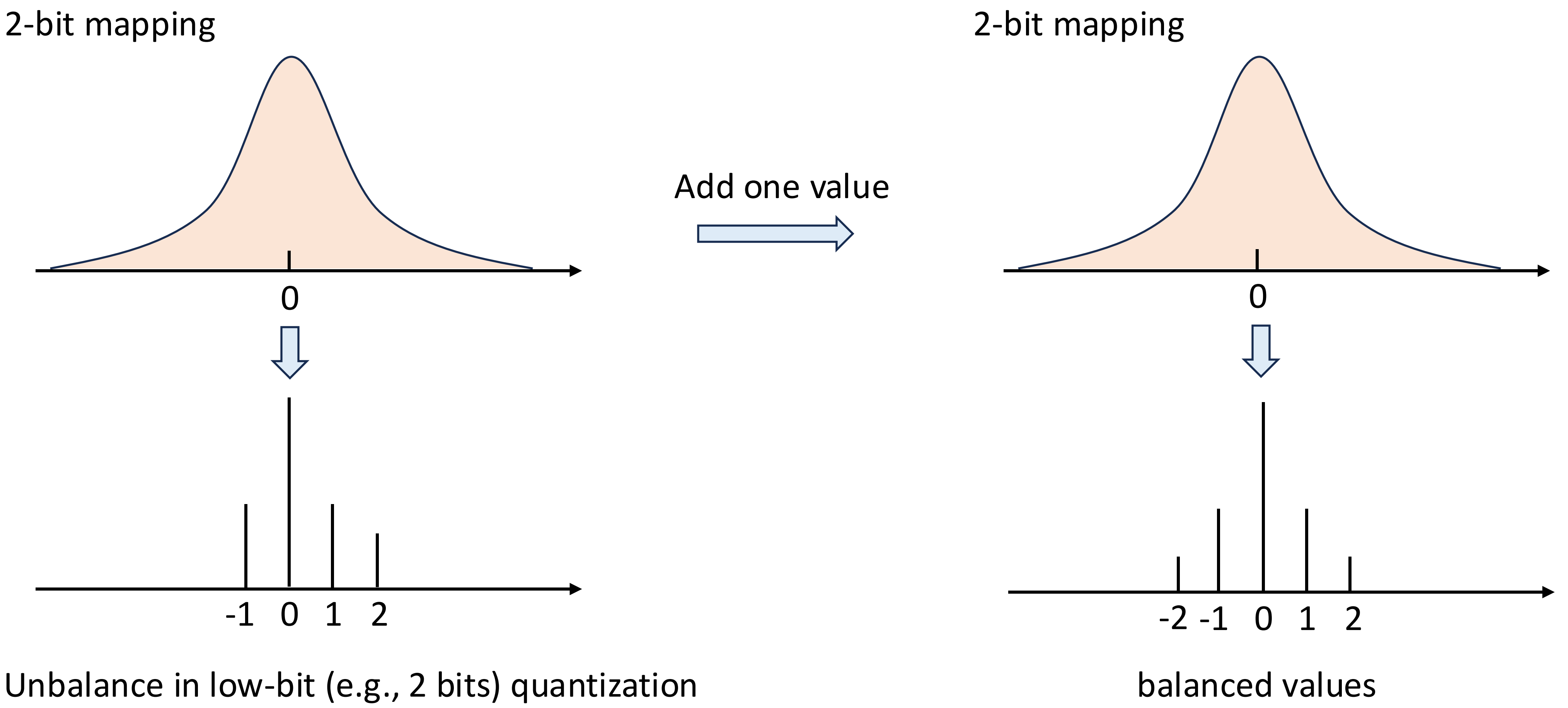
2-bit mapping



Unbalance in low-bit (e.g., 2 bits) quantization

Initialization

Adding Balance Integer

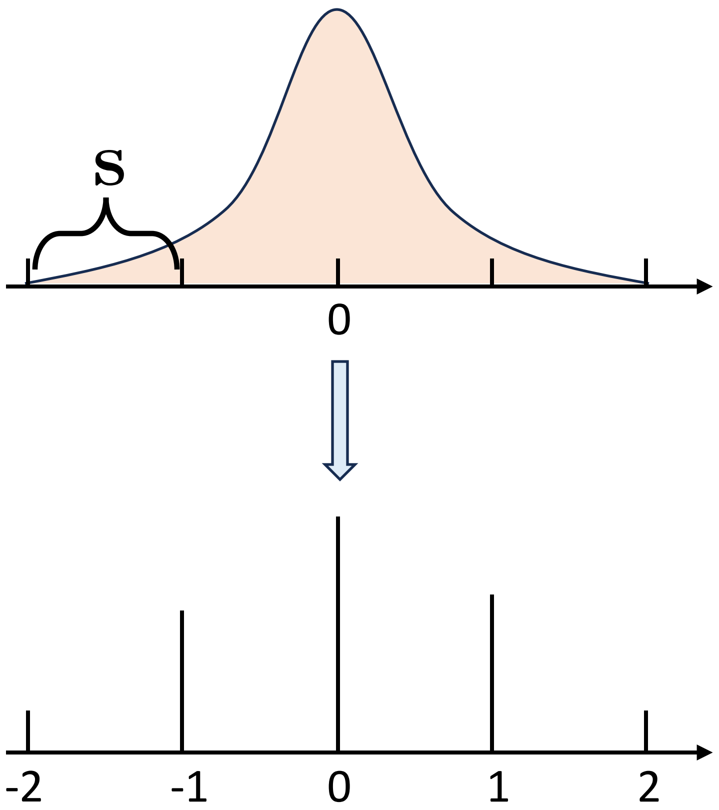


Initialization

Scaling Factor Initialization via Alternating Optimization

2-bit mapping

Min-Max initialization mapping



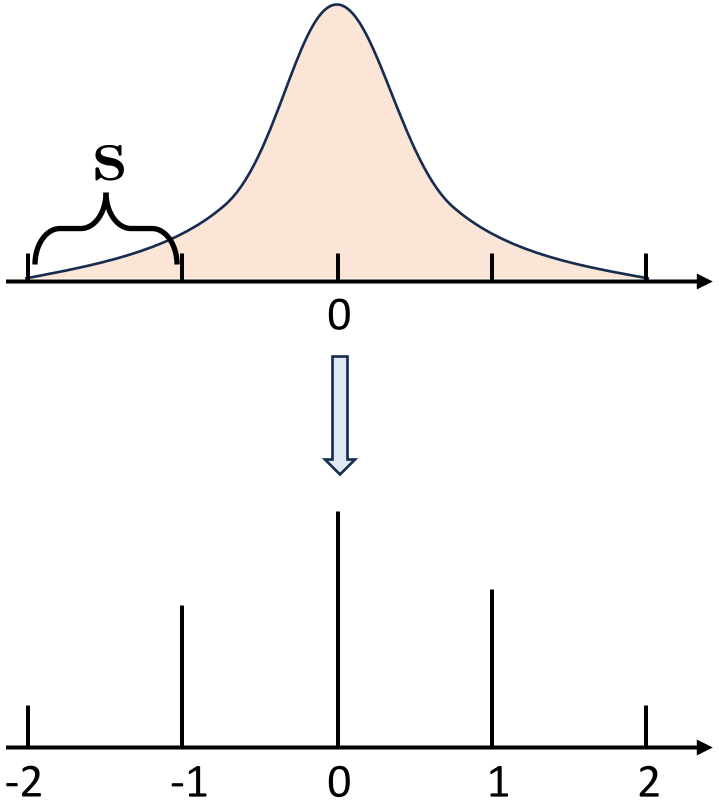
Drawback:
Large quantization error in
low-bit (e.g., 2 bits) quantization

Initialization

Scaling Factor Initialization via Alternating Optimization

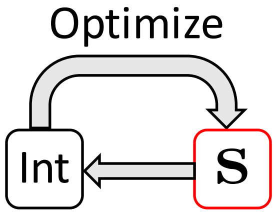
2-bit mapping

Min-Max initialization mapping



Drawback:
Large quantization error in
low-bit (e.g., 2 bits) quantization

Minimize Initial Quantization Error
By Updating Scaling Factor

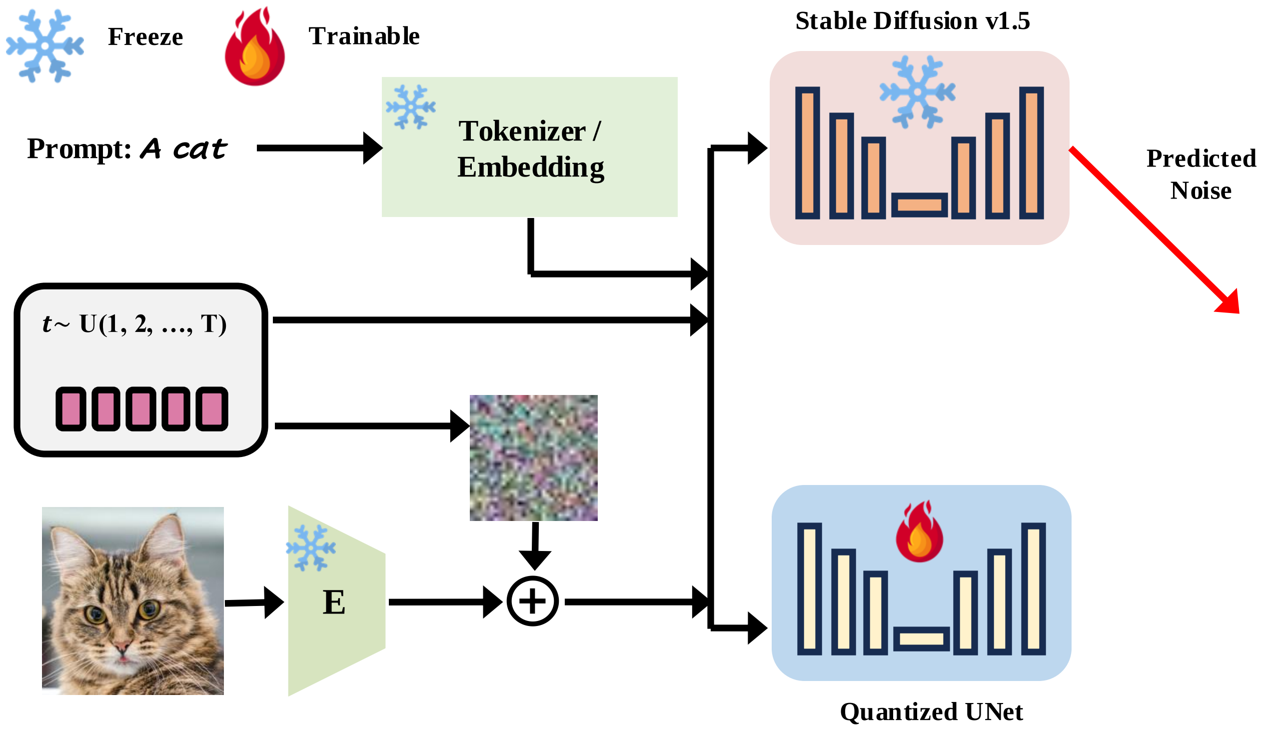


Two-stage Training Pipeline

Stage-I Training

Loss Function

CFG-aware Quantization Distillation

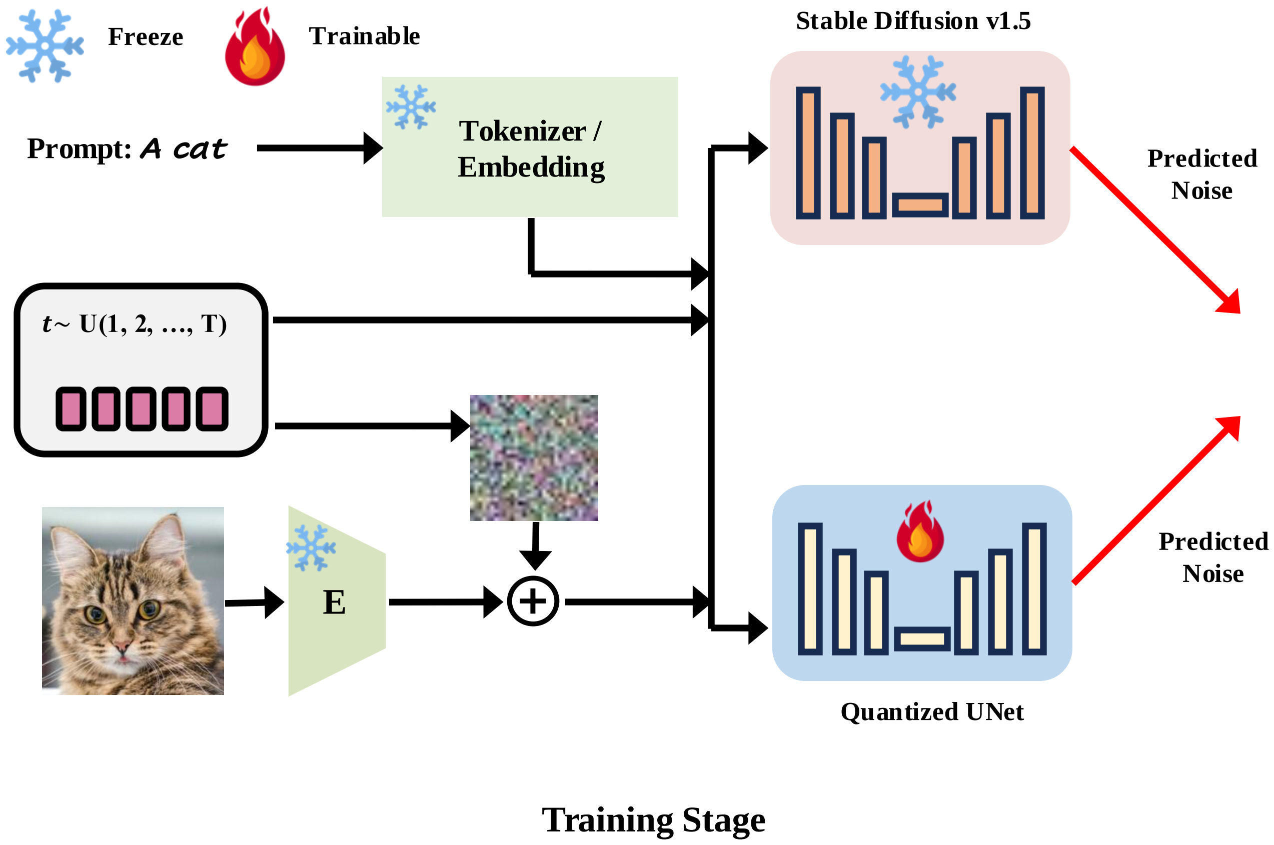


Training Stage

Stage-I Training

Loss Function

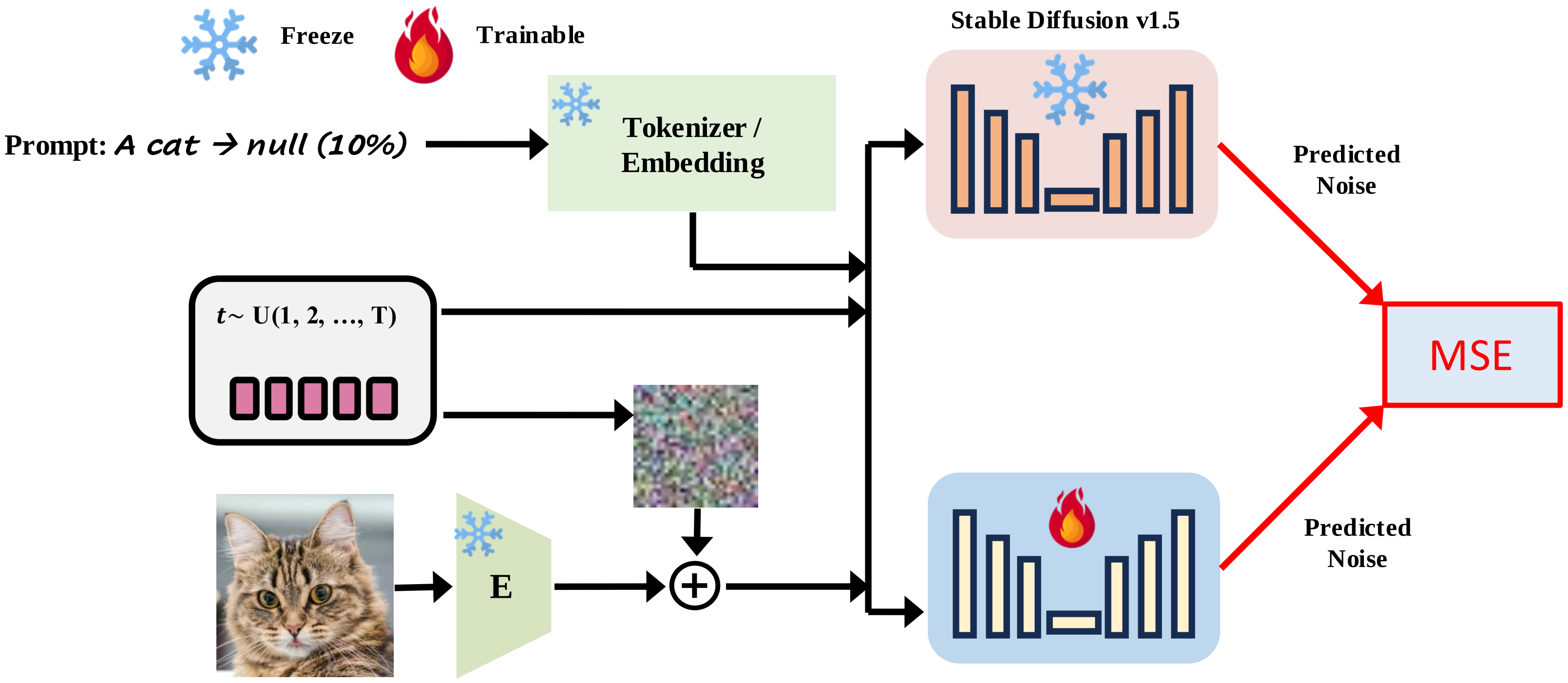
CFG-aware Quantization Distillation



Stage-I Training

Loss Function

CFG-aware Quantization Distillation

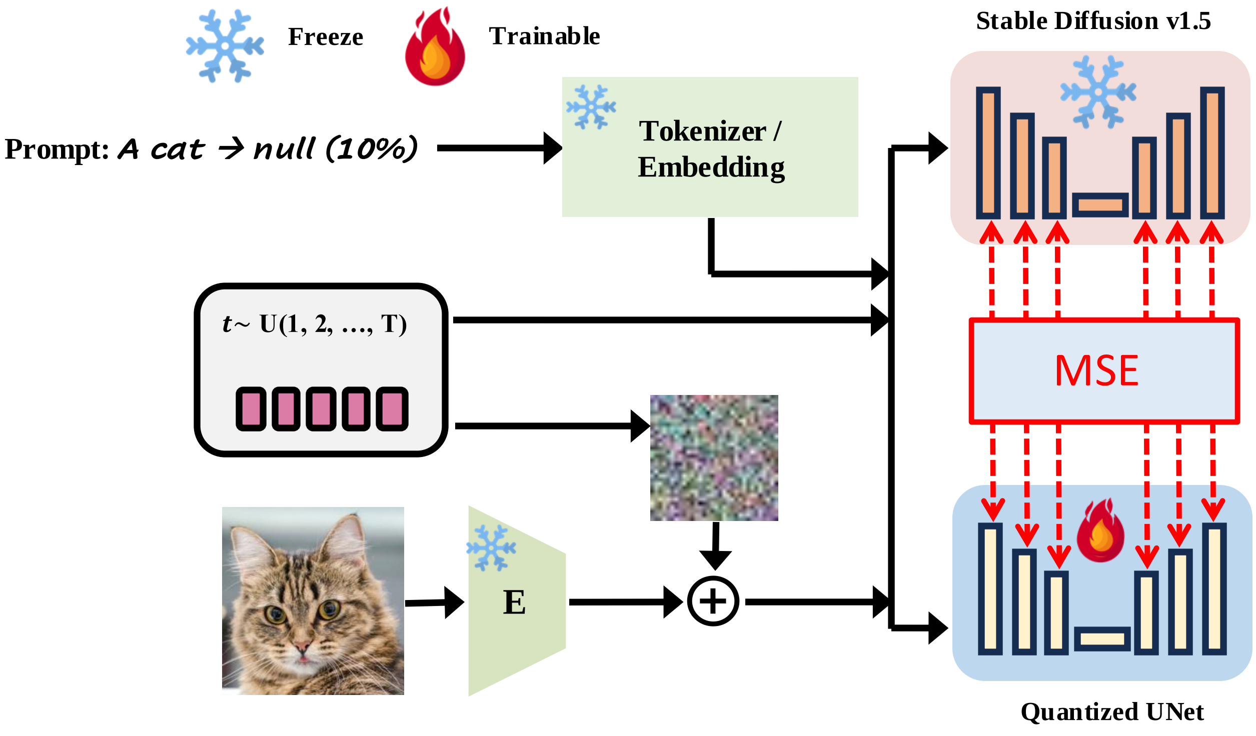


Training Stage

Stage-I Training

Loss Function

Feature Distillation

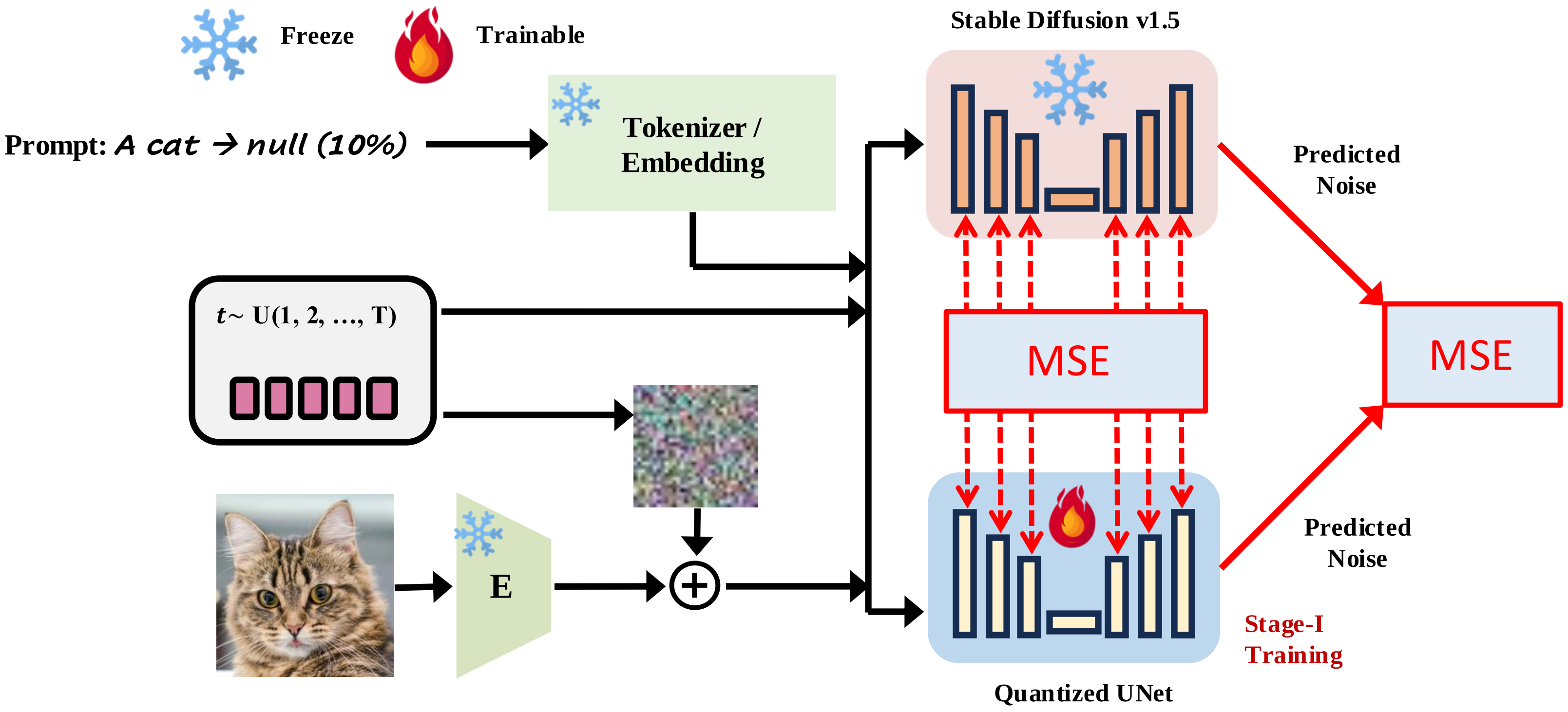


Training Stage

Stage-I Training

Loss Function

Overall Distillation



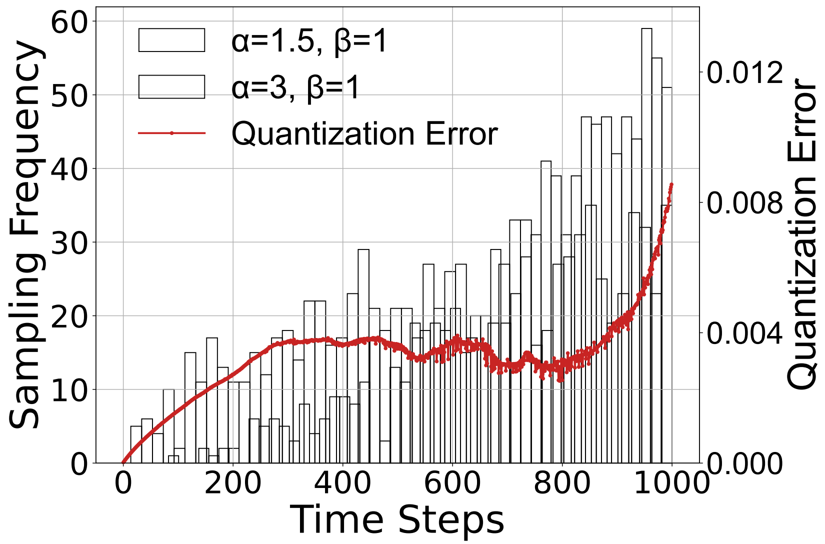
Training Stage

Stage-I Training

Quantization Error-aware Time Step Sampling

Motivation: Different Quantization Error at Different Time Steps

Quantization error of predicted latent features between quantized model and FP model



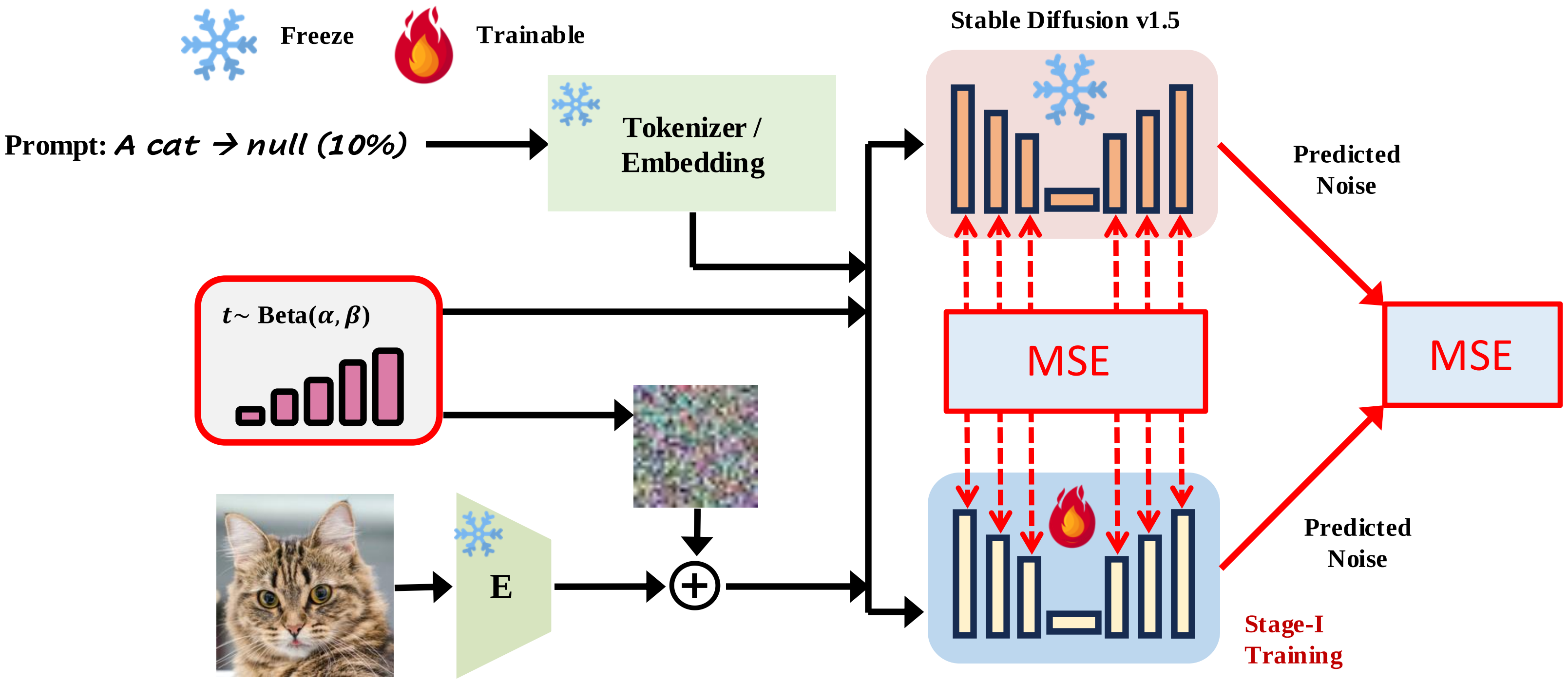
Observation: the quantization error keeps increasing as the time steps approach t = 999.

Solution: Sample more time steps exhibiting the larger quantization errors near t = 999 by Beta distribution.

Stage-I Training

Loss Function

Overall Distillation

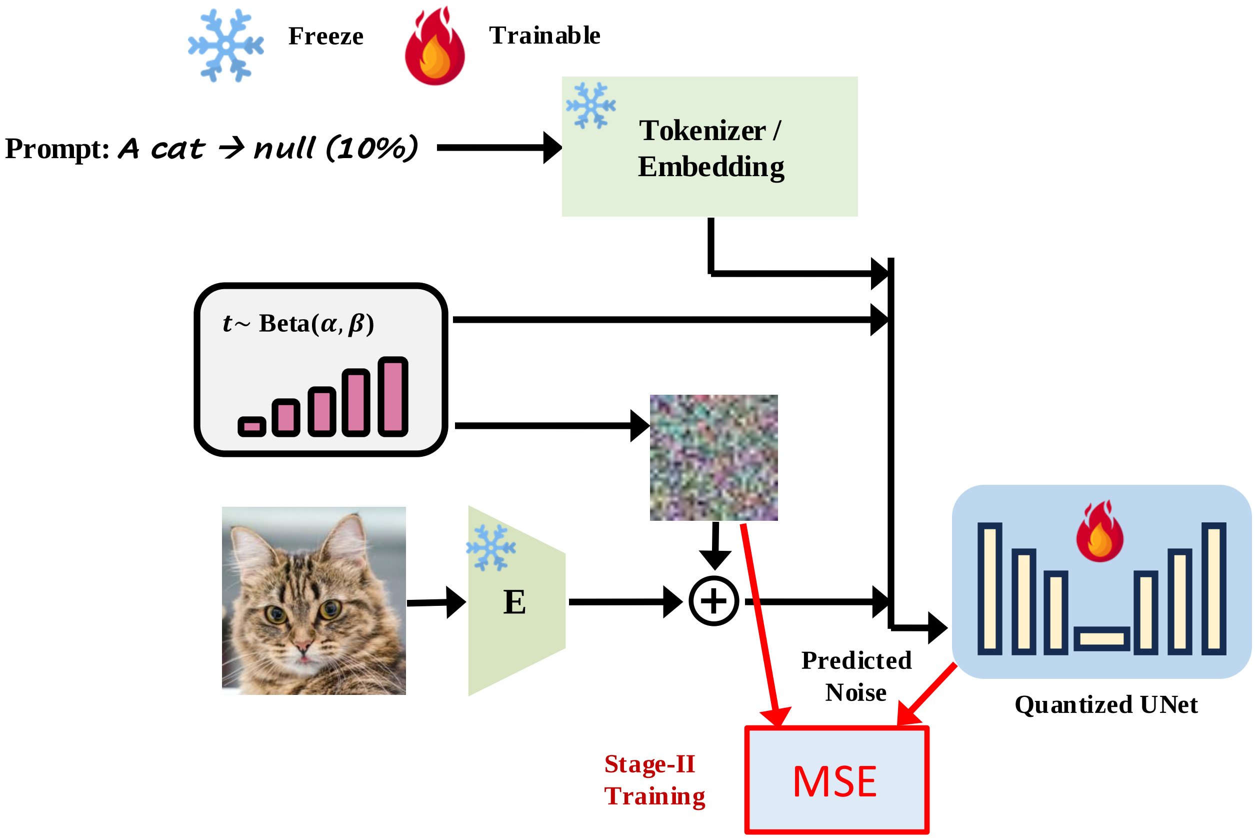


Training Stage

Stage-II Training

Fine-tuning with Noise Prediction

Stage-II Loss



Training Stage

Results

Results

Generated Images

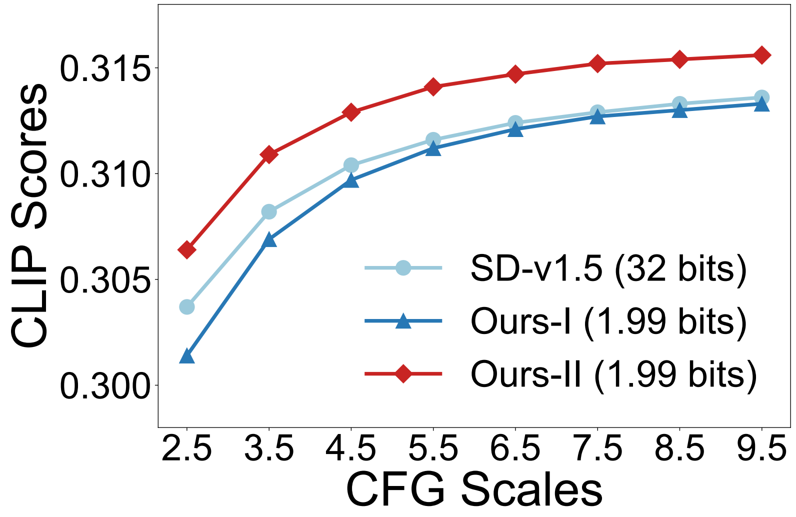
Sampler: PNDM
Steps: 50
Seed: 1024

Stable Diffusion v1.5, 32 bits

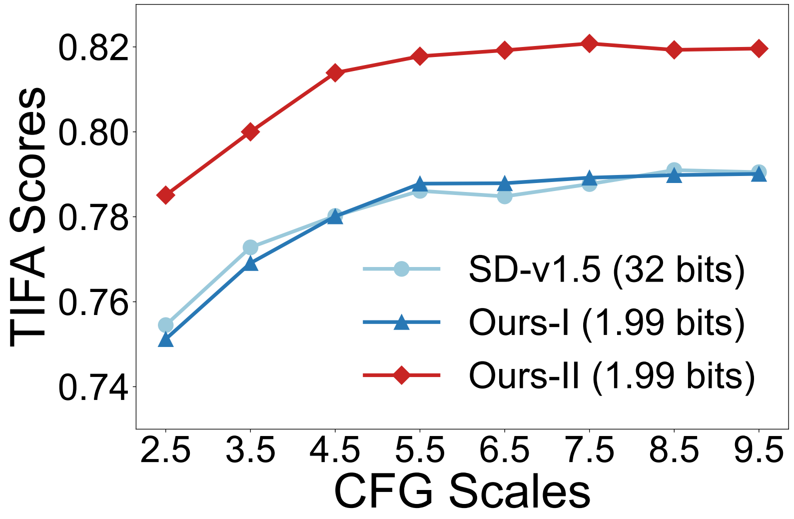
BitsFusion, 1.99 bits

Results

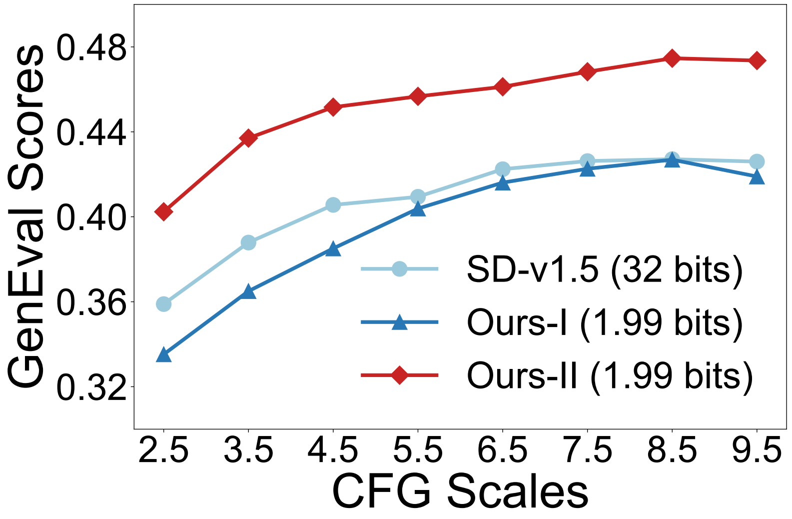
Quantitative performance



CLIP Score on 30K MS-COCO.



TIFA Scores



GenEval Scores

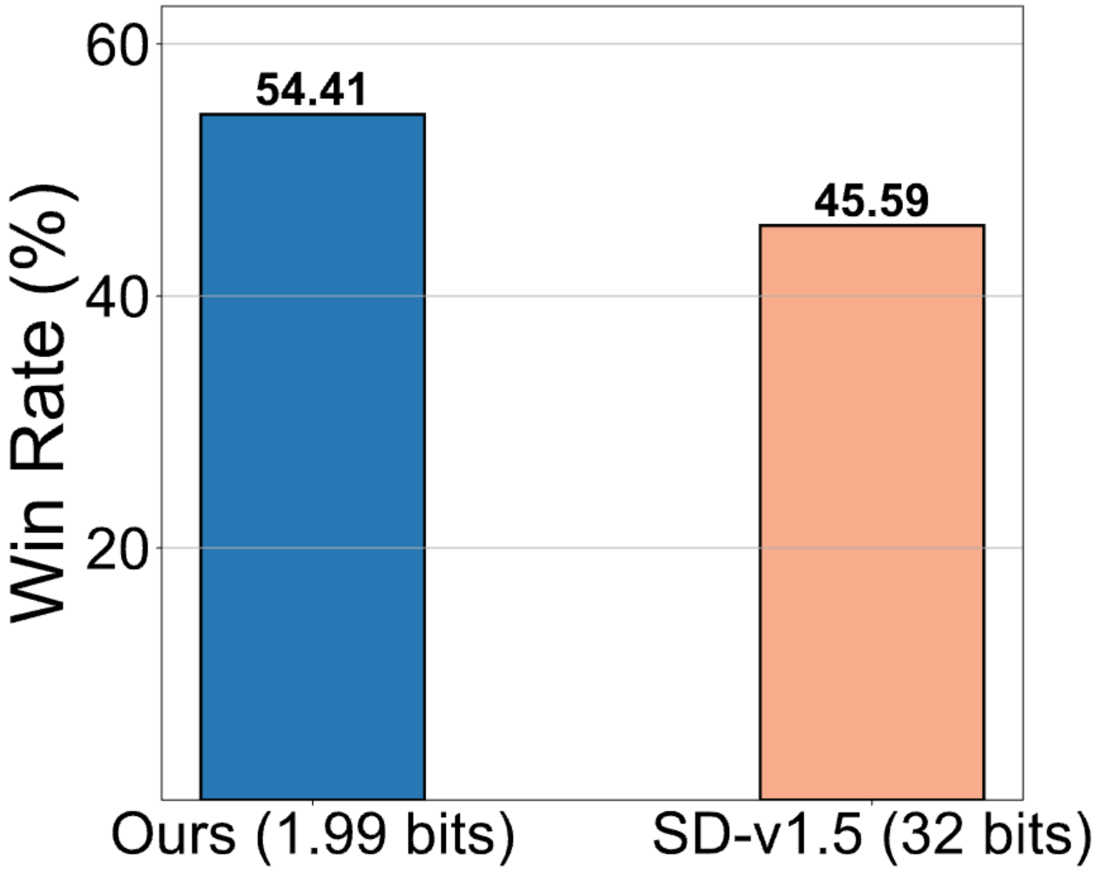
Ours-I: Stage-I training
Ours-II: Stage-II training

BitsFusion consistently outperforms Stable Diffusion v1.5

Results

Human Evaluation

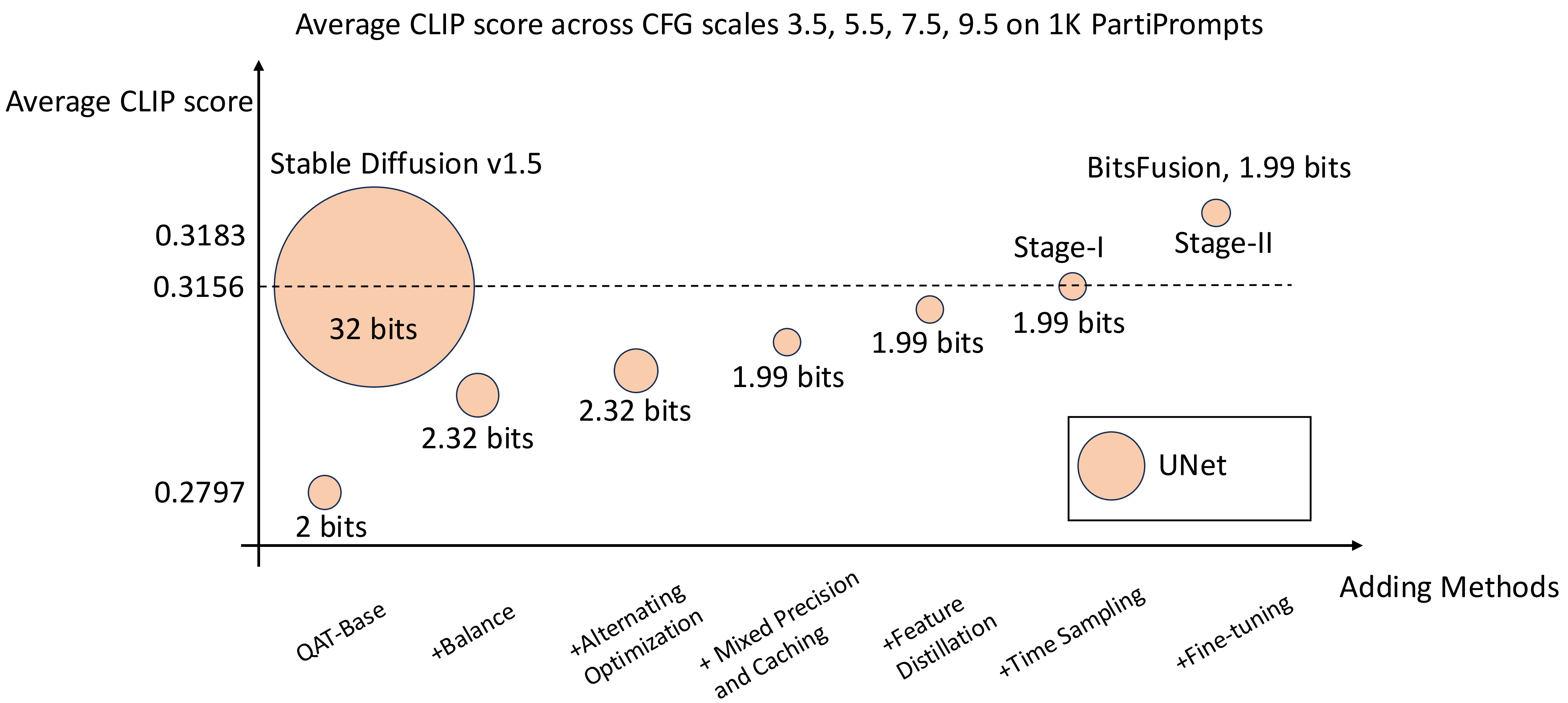
Given a prompt, which image has better aesthetics and image-text alignment?



User preference of generated images from PartiPrompts (P2)

Results

Effect of each method



Results

More comparisons

Sampler: PNDM
Steps: 50
Seed: 1024

Stable Diffusion v1.5, 32 bits

BitsFusion, 1.99 bits

Results

More comparisons

Sampler: PNDM
Steps: 50
Seed: 1024

Stable Diffusion v1.5, 32 bits

BitsFusion, 1.99 bits

Results

More comparisons

Sampler: PNDM
Steps: 50
Seed: 1024

Stable Diffusion v1.5, 32 bits

BitsFusion, 1.99 bits

Results

More comparisons

Sampler: PNDM
Steps: 50
Seed: 1024

Stable Diffusion v1.5, 32 bits

BitsFusion, 1.99 bits

Results

More comparisons

Sampler: PNDM
Steps: 50
Seed: 1024

Stable Diffusion v1.5, 32 bits

BitsFusion, 1.99 bits

Results

More comparisons

Sampler: PNDM
Steps: 50
Seed: 1024

Stable Diffusion v1.5, 32 bits

BitsFusion, 1.99 bits

Results

More comparisons

Sampler: PNDM
Steps: 50
Seed: 1024

Stable Diffusion v1.5, 32 bits

BitsFusion, 1.99 bits

Thank you