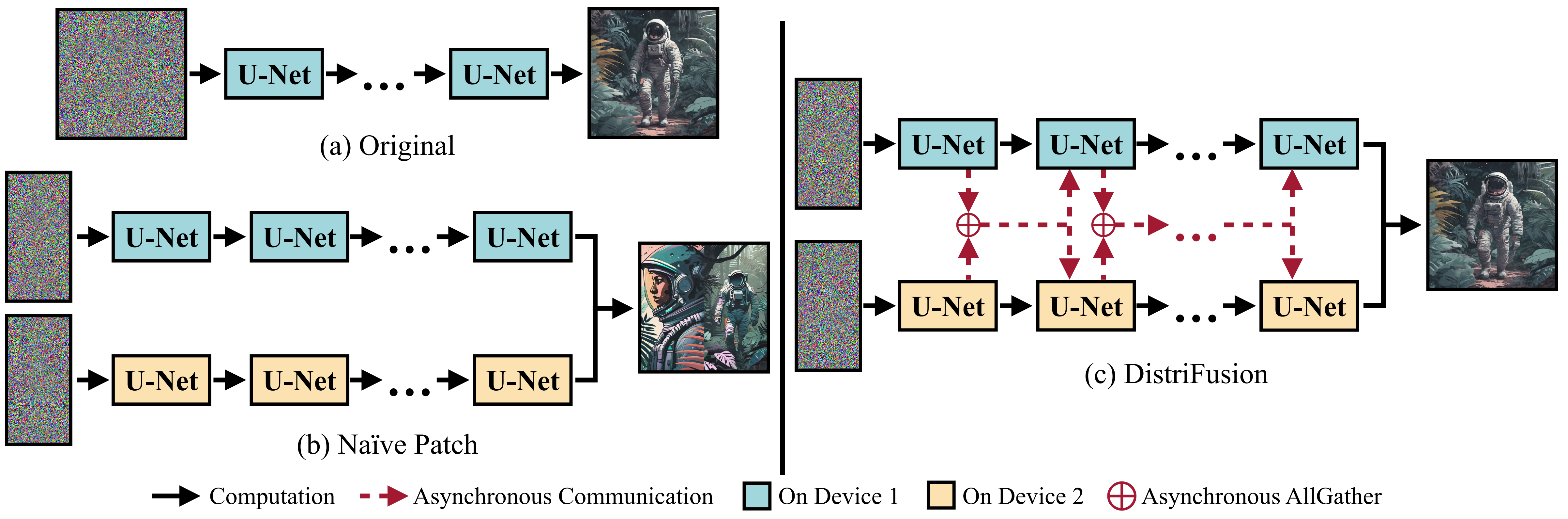
# DistriFusion: Distributed Parallel Inference for High-Resolution Diffusion Models

Muyang Li[1*], Tianle Cai[2*], Jiaxin Cao[3], Qinsheng Zhang[4], Han Cai[1], Junjie Bai[3], Yangqing Jia[3], Ming-Yu Liu[4], Kai Li[3] and Song Han[1,4]
[1]MIT [2]Princeton [3]Lepton AI [4]NVIDIA

```
pip install distrifuser
```

## Introduction



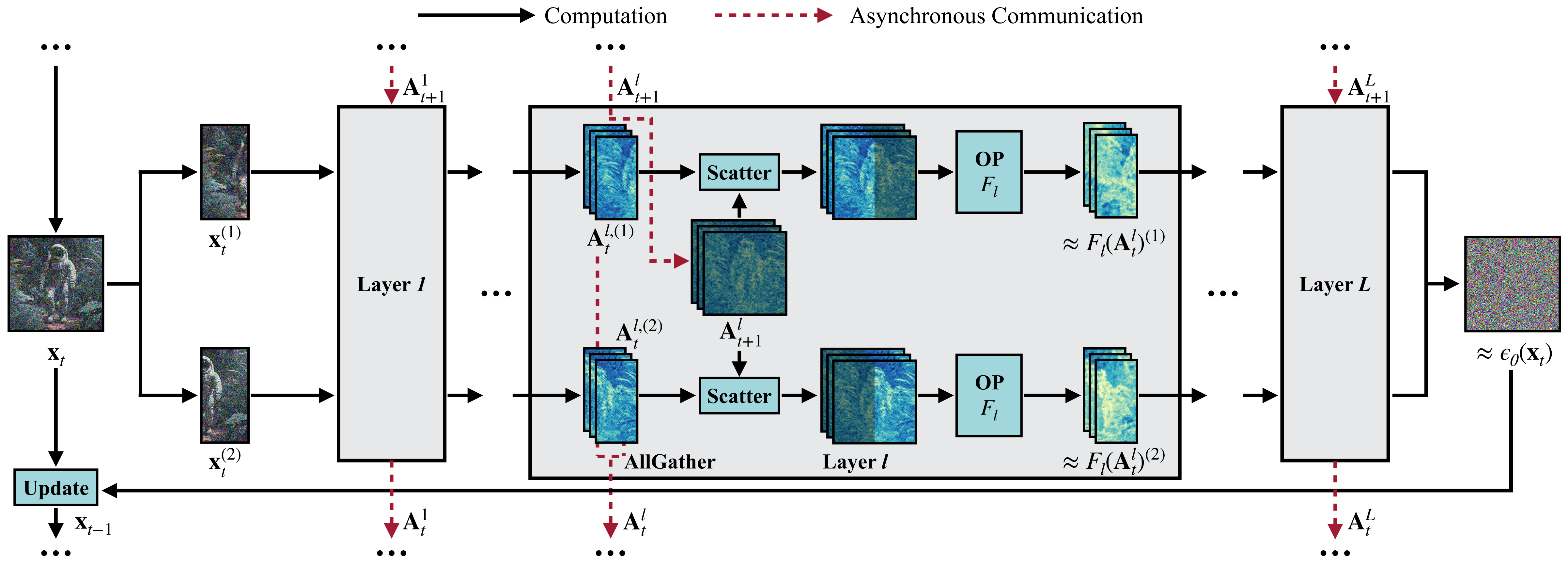
### Background:

- Leverage multiple GPUs to accelerate diffusion models.
- Naïve Patch has artifacts due to lack of patch interaction.
- Introducing interactions will incur communication overheads.

### Solution — DistriFusion:

- Distribute image patches across GPUs.
- Use the similarity between inputs of adjacent timesteps.
- Reuse the previous features.
- Hide communication costs with asynchronous communication.

## Overview



### Displace Patch Parallelism:

- Split the image into patches for each device.
- Use async AllGather to cache features for the next step.
- Scatter the fresh activation into the previous features.
- Only perform computation at the fresh regions.

### Sparse Operations:

- Conv — Apply kernel to the fresh regions.
- Attn — Fresh regions attend to entire scattered features.

### Other Optimizations:

- Corrected asynchronous GN to avoid GN synchronization.
- Adding warm-up steps to improve the quality.

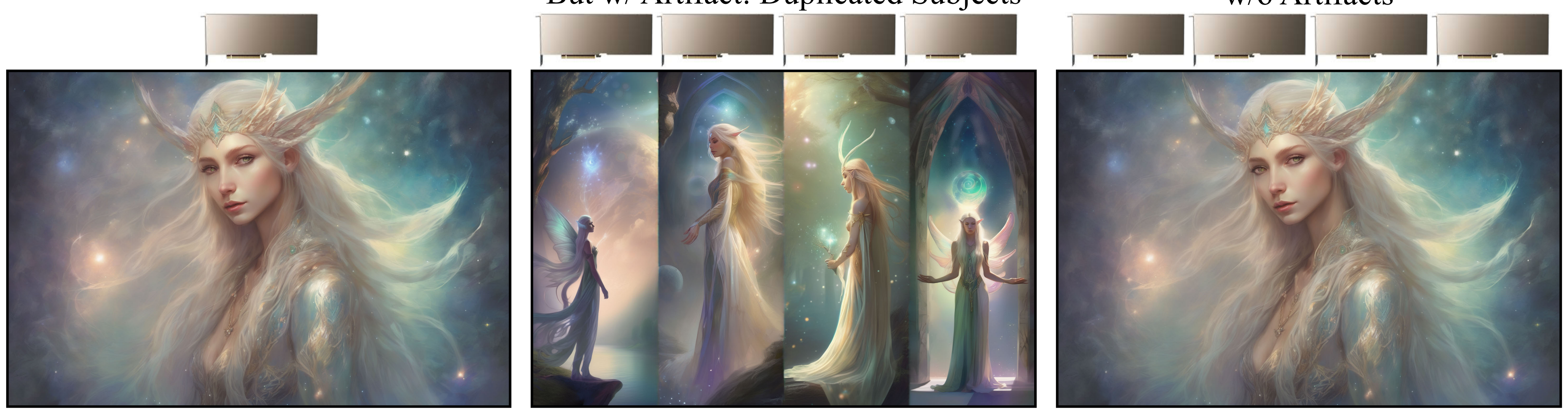
## Results

### Quality Results:

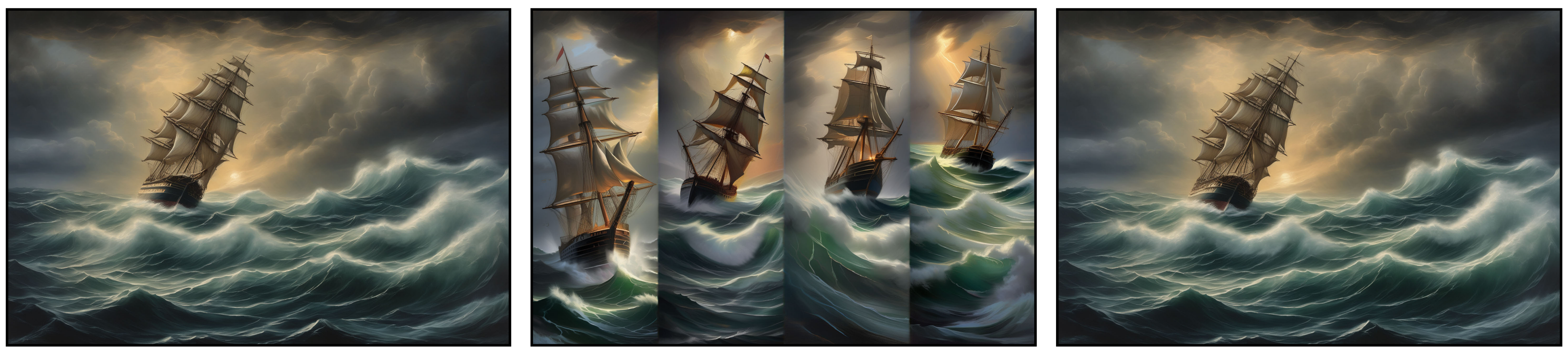
Original, 1 GPU
MACs: 907T
Latency: **12.3s**

Naïve Parallelization, 4 GPUs
MACs Per Device: 190T (4.8× Less)
Latency: 3.14s (3.9× Faster)
But w/ Artifact: Duplicated Subjects

DistriFusion (Ours), 4 GPUs
MACs Per Device: 227T **(4.0× Less)**
Latency: **4.16s (3.0× Faster)**
w/o Artifacts

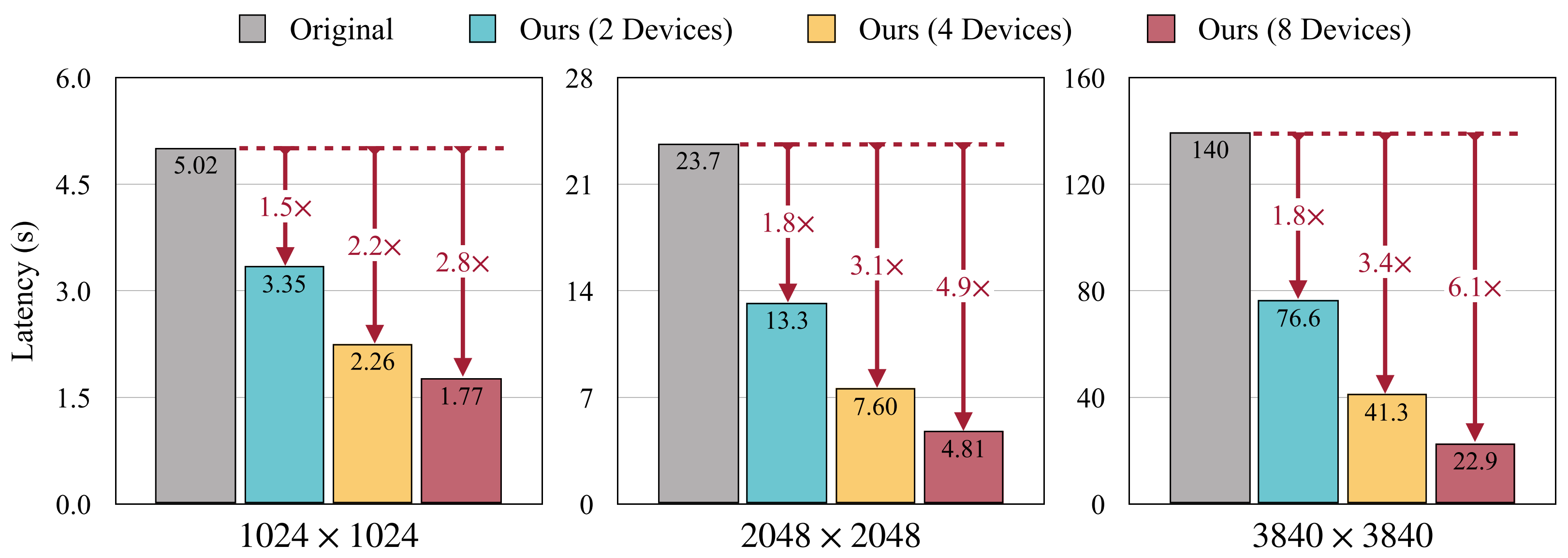


Prompt: *Ethereal fantasy concept art of an elf, magnificent, celestial, ethereal, painterly, epic, majestic, magical, fantasy art, cover art, dreamy.*



Prompt: *Romantic painting of a ship sailing in a stormy sea, with dramatic lighting and powerful waves.*

### Speedups:



### Compare to Tensor Parallelism:

| Method | 1024 × 1024 | | 2048 × 2048 | | 3840 × 3840 | |
|---|---|---|---|---|---|---|
| | Comm. | Latency | Comm. | Latency | Comm. | Latency |
| Original | – | 5.02s | – | 23.7s | – | 140s |
| Sync. TP | 1.33G | 3.61s | 5.33G | 11.7s | 18.7G | 46.3s |
| Sync. PP | 0.42G | 2.21s | 1.48G | 5.62s | 5.38G | 24.7s |
| **DistriFusion (Ours)** | **0.42G** | **1.77s** | **1.48G** | **4.81s** | **5.38G** | **22.9s** |
| No Comm. | – | 1.48s | – | 4.14s | – | 21.3s |

### More Visualization:

Original
Latency: 5.02s
FID: 24.0

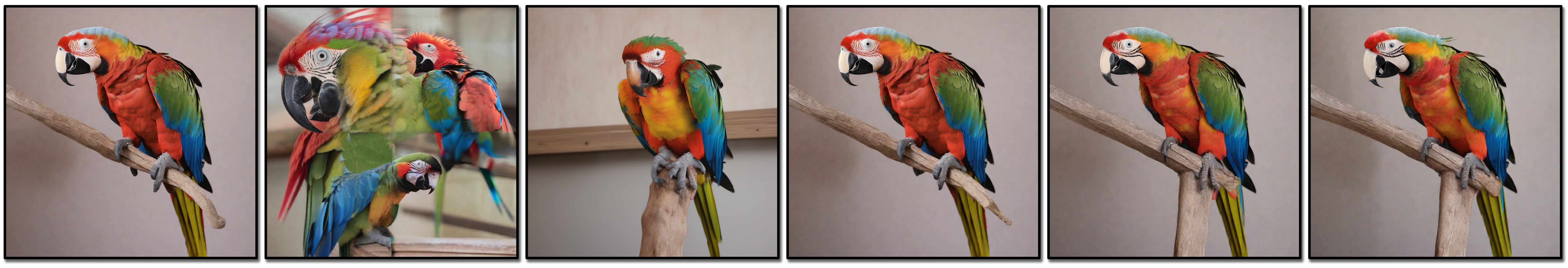
Naïve Patch (2 Devices)
Latency: 2.83s (1.8× Faster)
FID: 33.6

ParaDiGMS (8 Devices)
Latency: 1.80s (2.8× Faster)
FID: 25.1

Ours (2 Devices)
**Latency: 3.35s (1.5× Faster)**
**FID: 24.0**

Ours (4 Devices)
**Latency: 2.26s (2.2× Faster)**
**FID: 24.2**

Ours (8 Devices)
**Latency: 1.77s (2.8× Faster)**
**FID: 24.3**



Prompt: *A multi-colored parrot holding its foot up to its beak.*

Hardware Accelerated Neural-nets

PRINCETON UNIVERSITY

NVIDIA

CVPR
JUNE 17-21, 2024
SEATTLE, WA