NeuRex: A Case for **Neu**ral **Re**ndering **Acc**eleration

Junseo Lee   Kwanseok Choi   Jungi Lee
Seokwon Lee   Joonho Whangbo   Jaewoong Sim

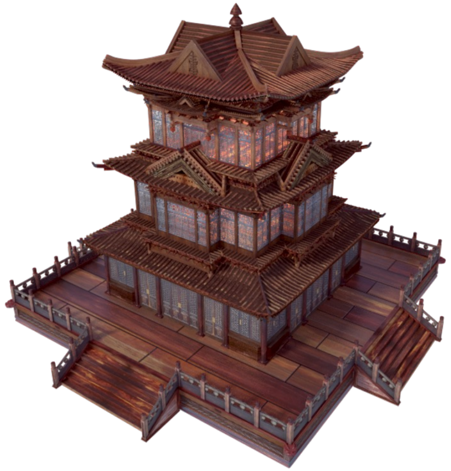
Seoul National University

ISCA-50 | June 2023

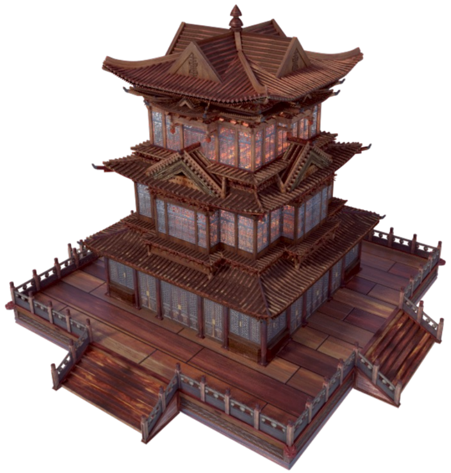
Department of ECE
Seoul National University

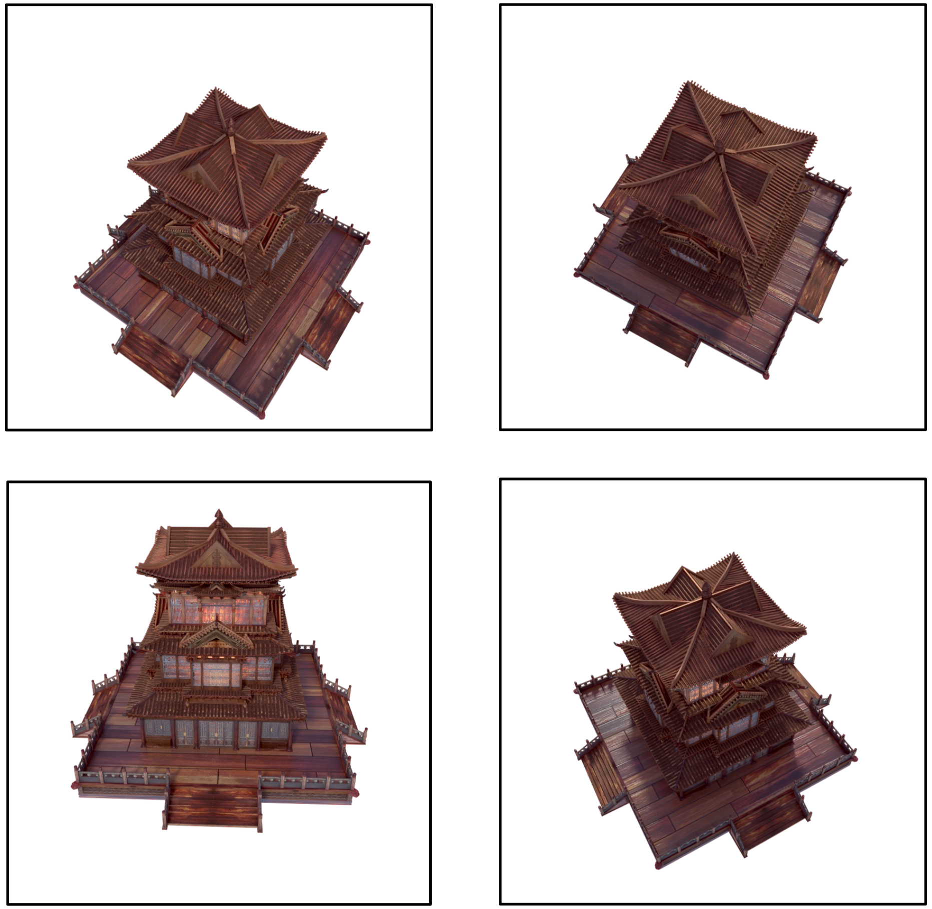
Neural Rendering

Neural Rendering

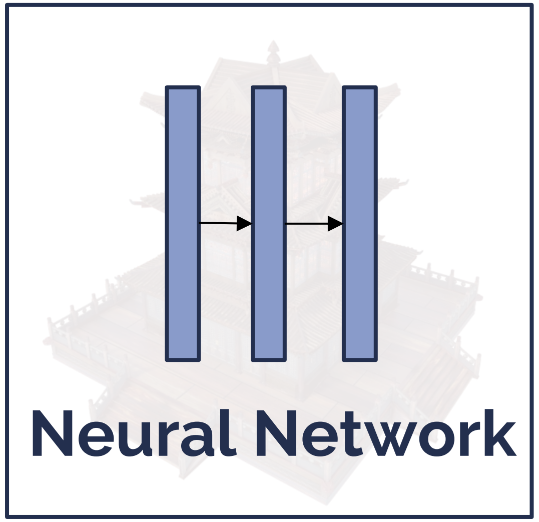


Neural Rendering

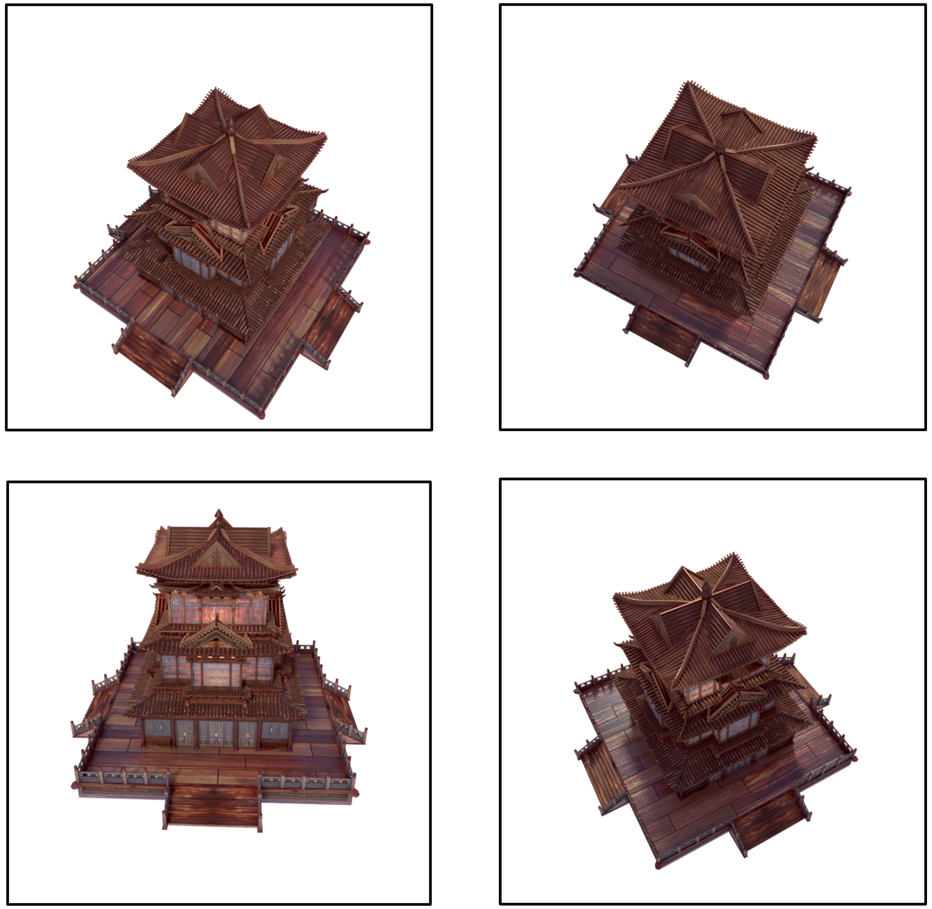




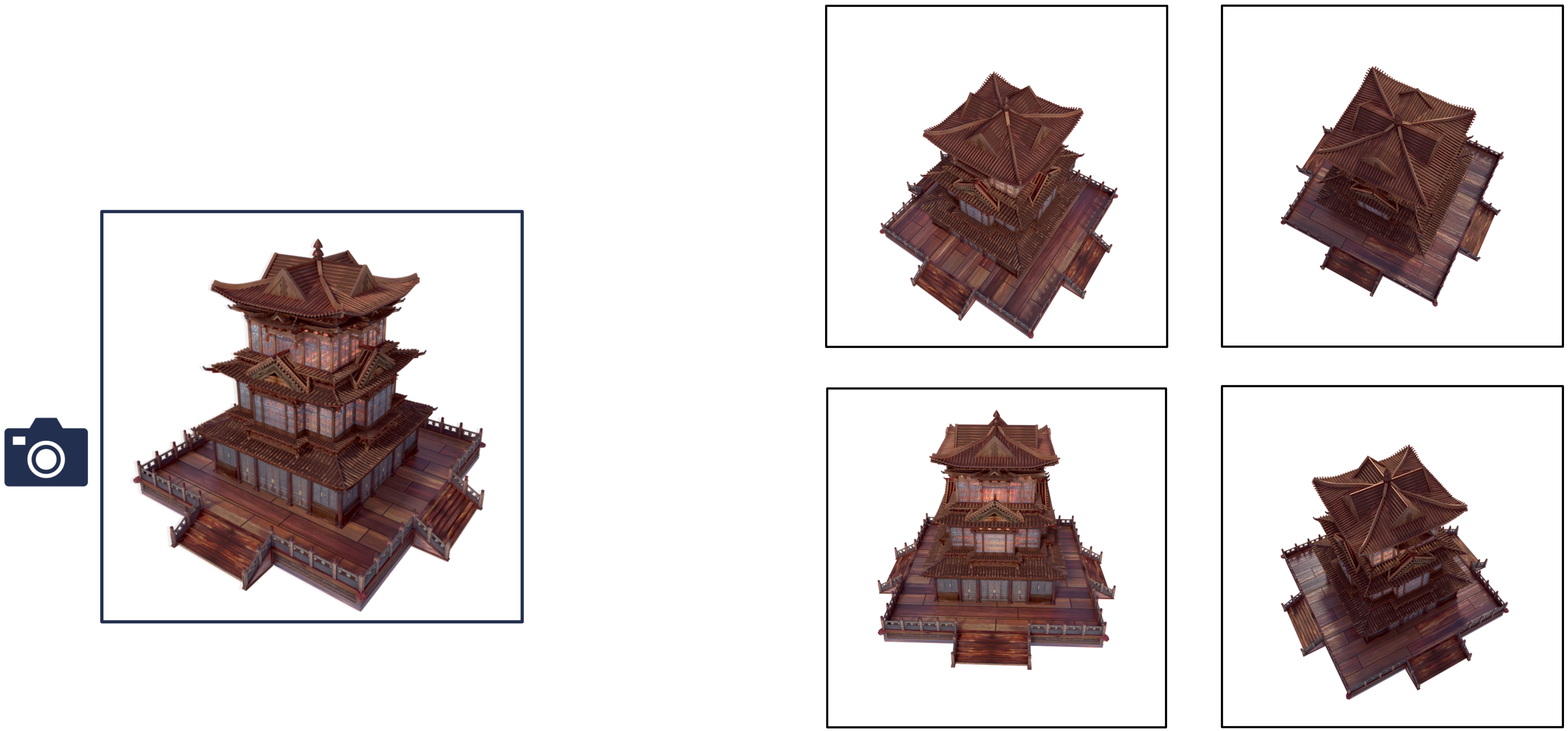
Neural Rendering



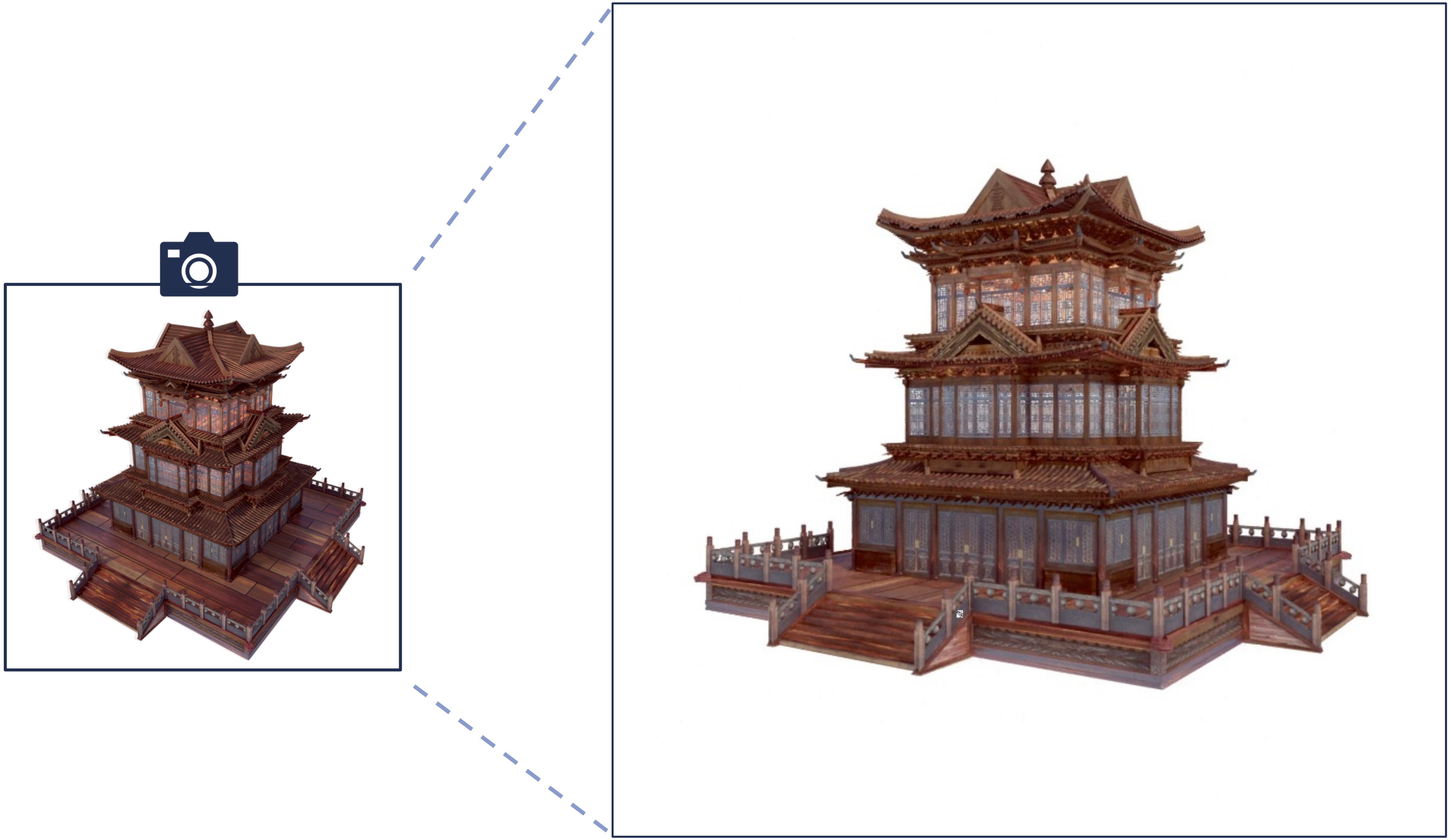
Neural Network



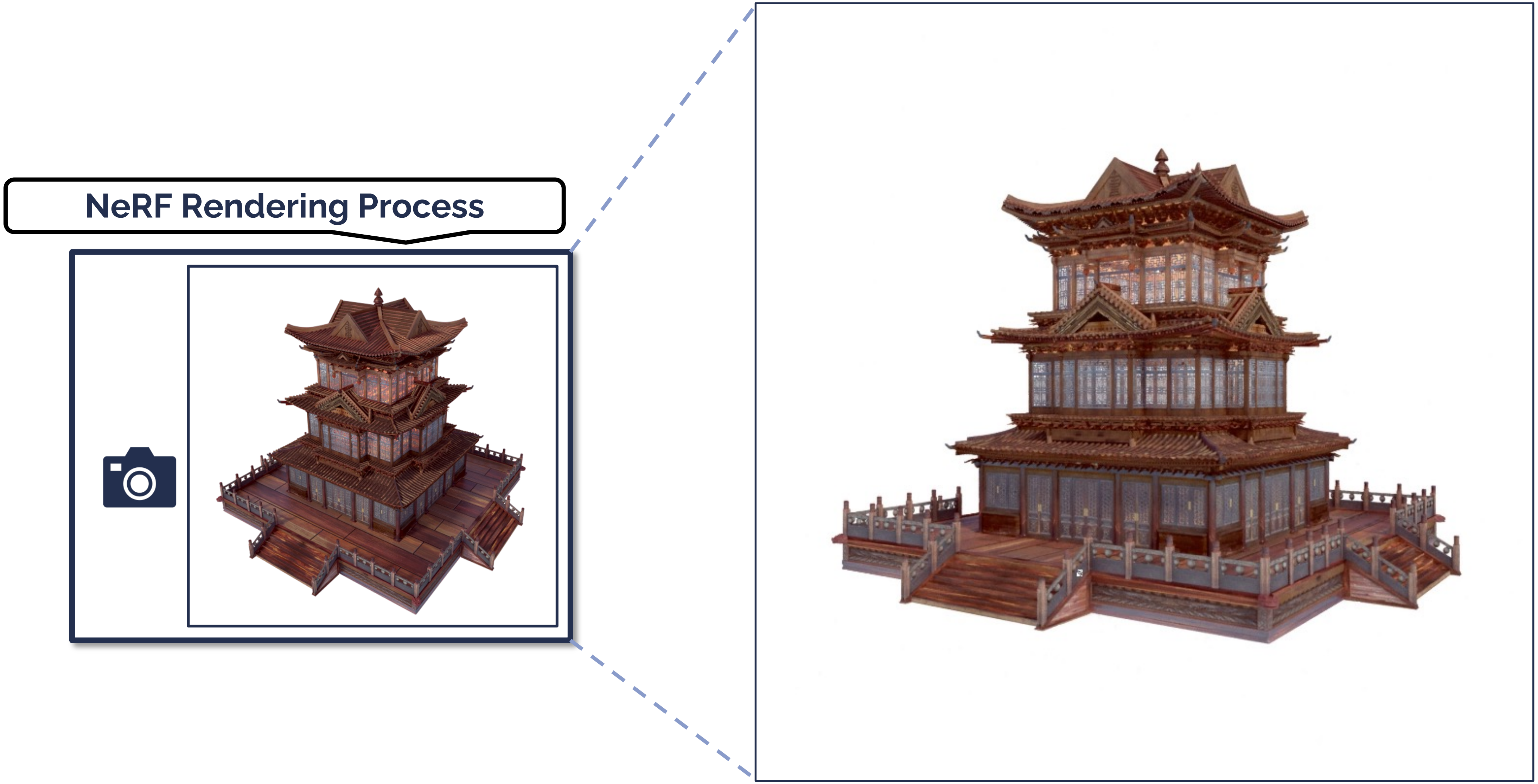
Neural Rendering



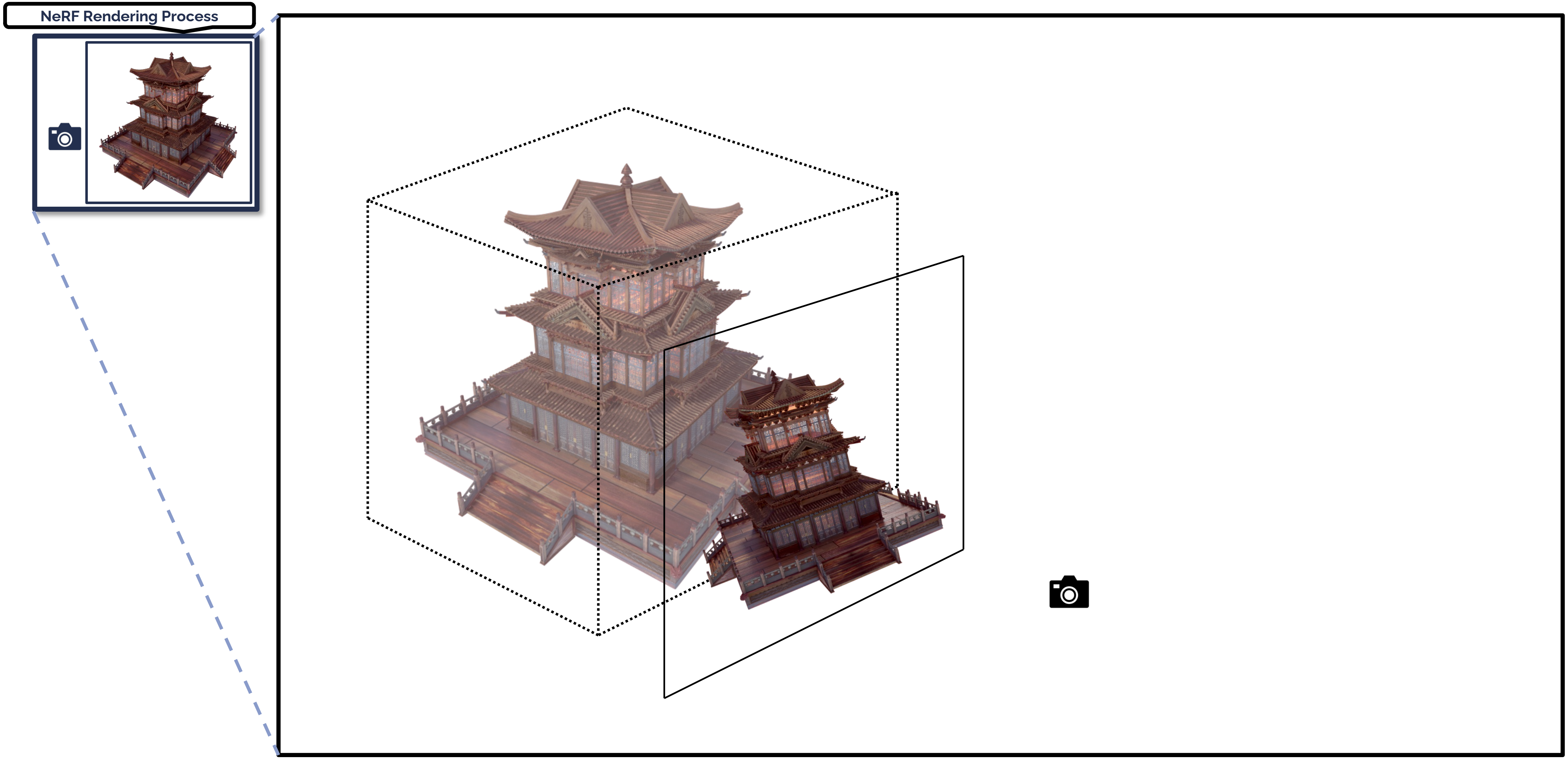
Neural Rendering



Neural Rendering

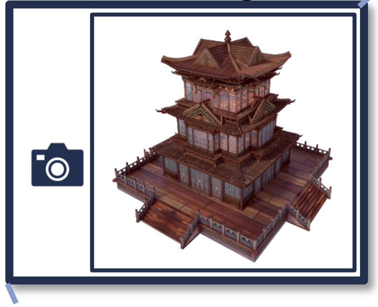

NeRF Rendering Process

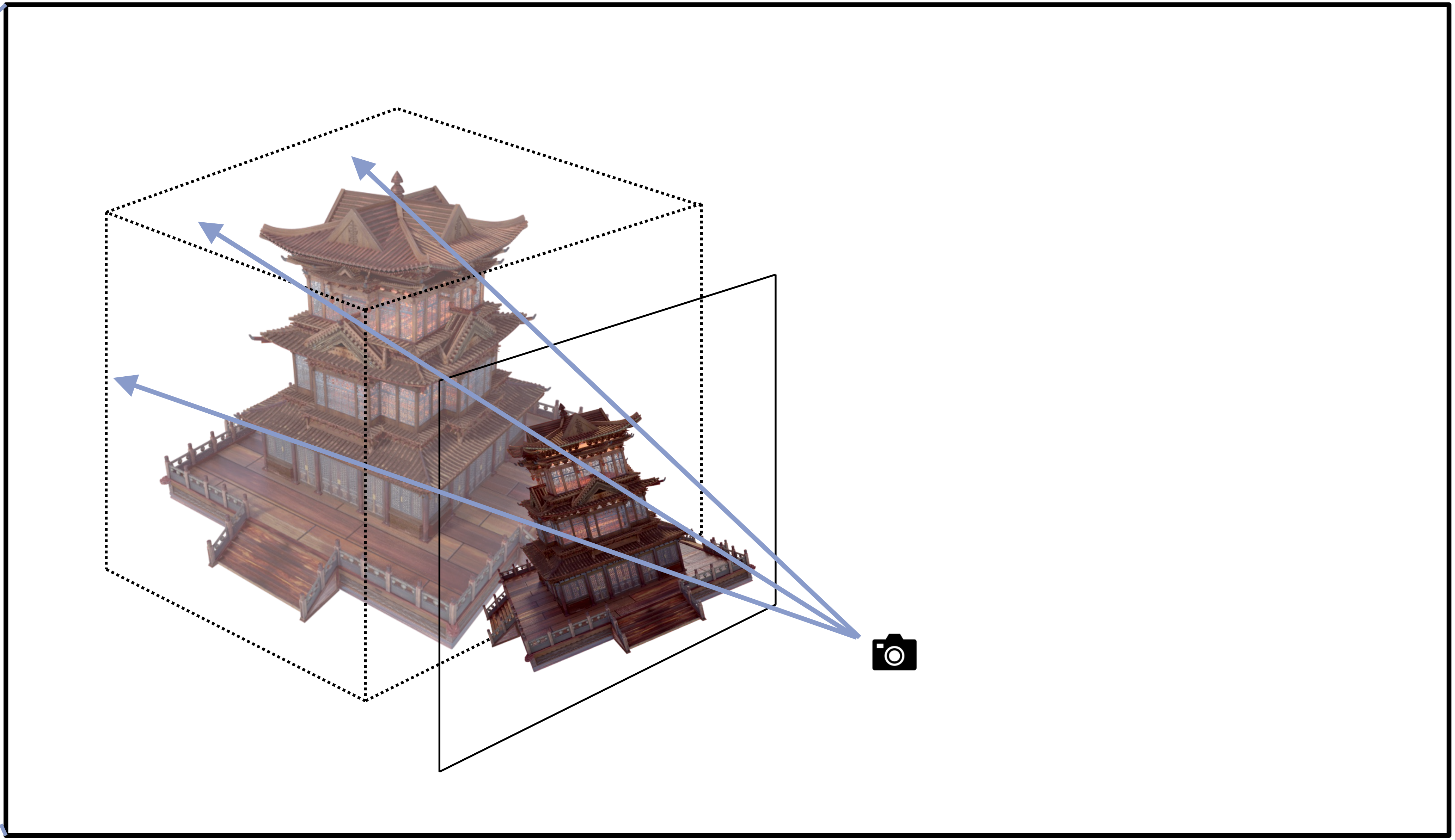
NeRF Rendering Process



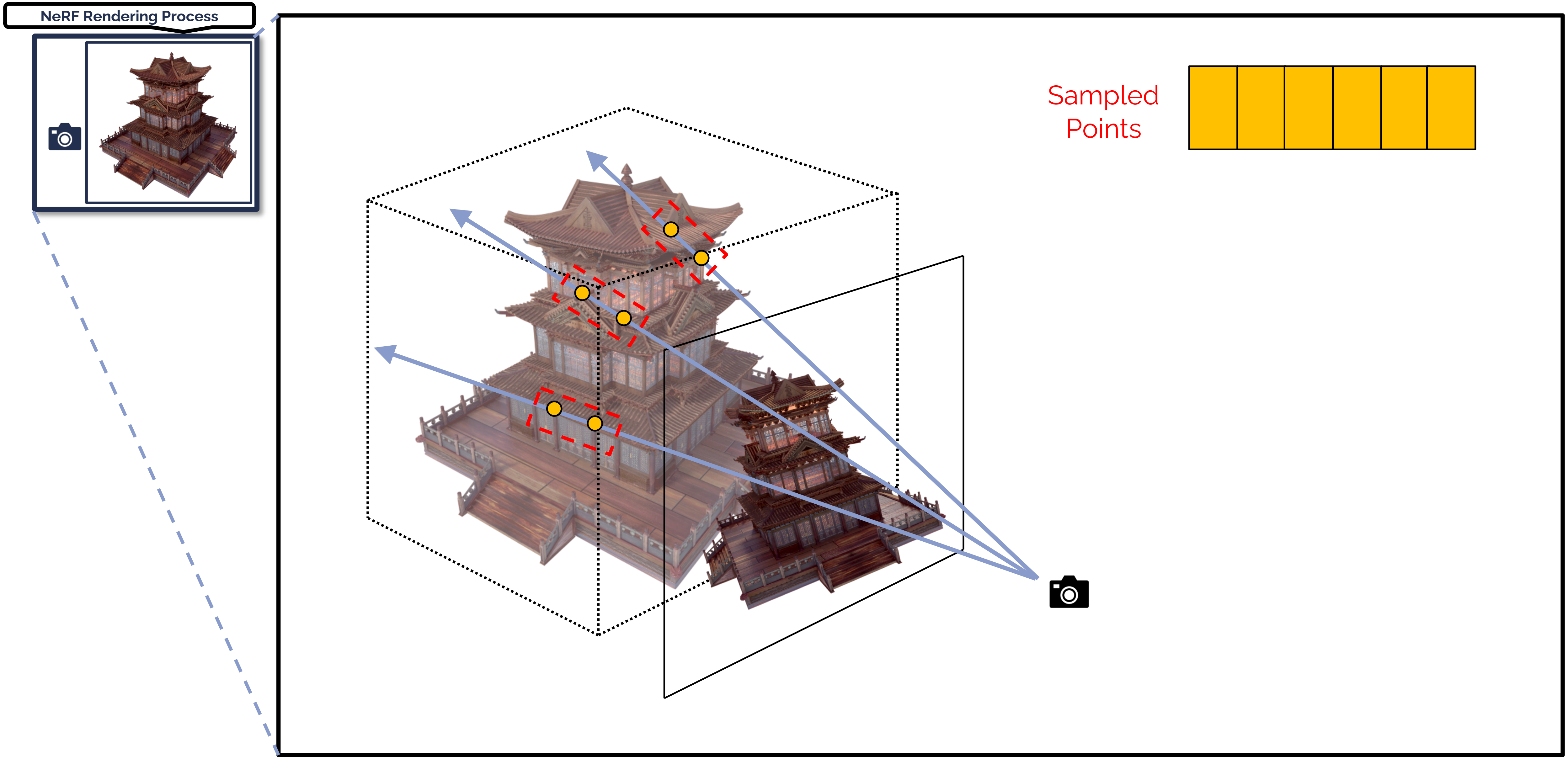
NeRF Rendering Process

NeRF Rendering Process



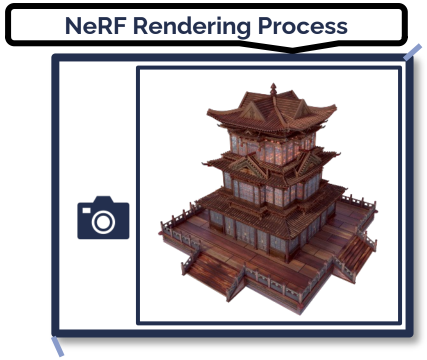


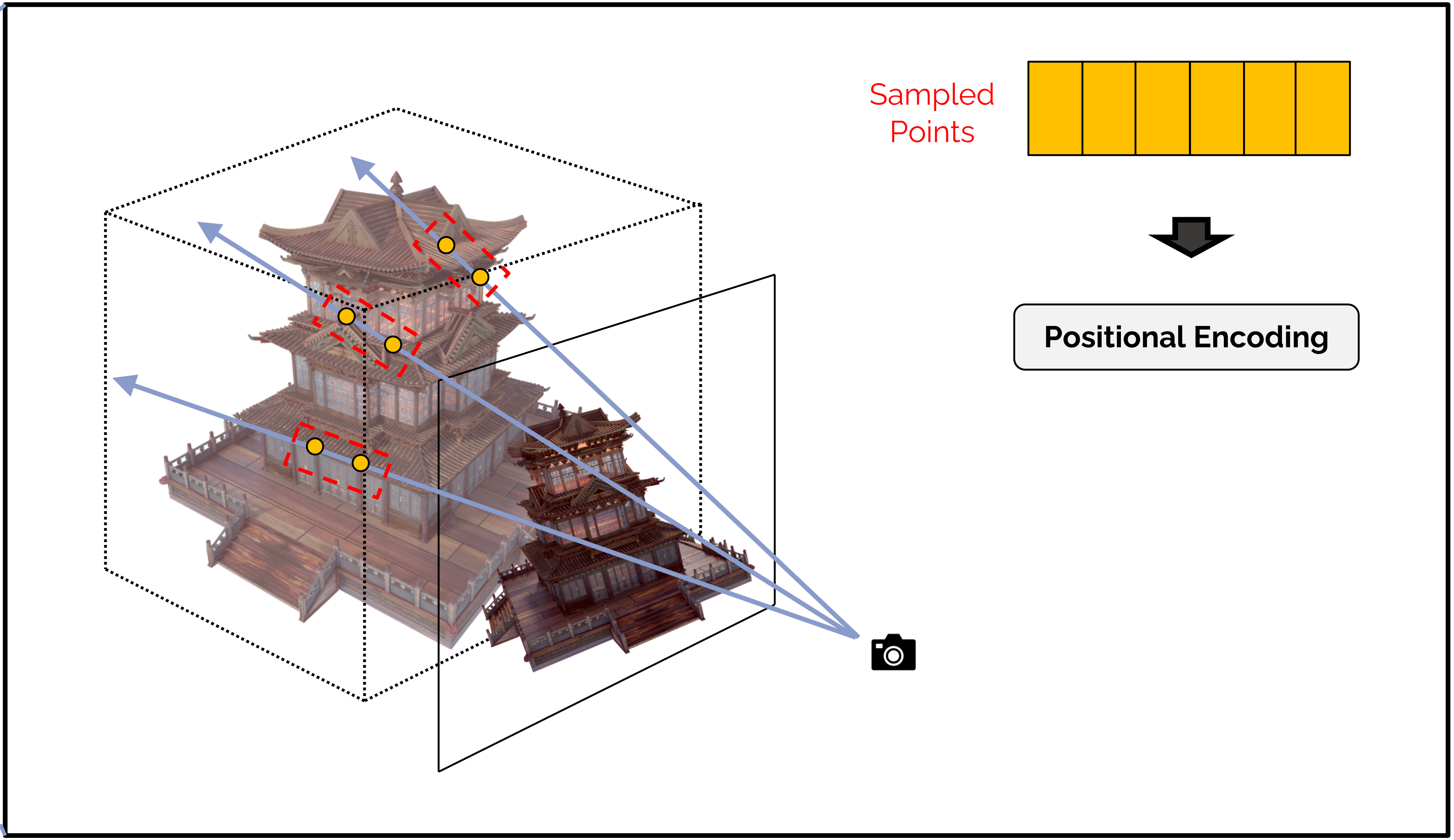
NeRF Rendering Process



Sampled Points

NeRF Rendering Process

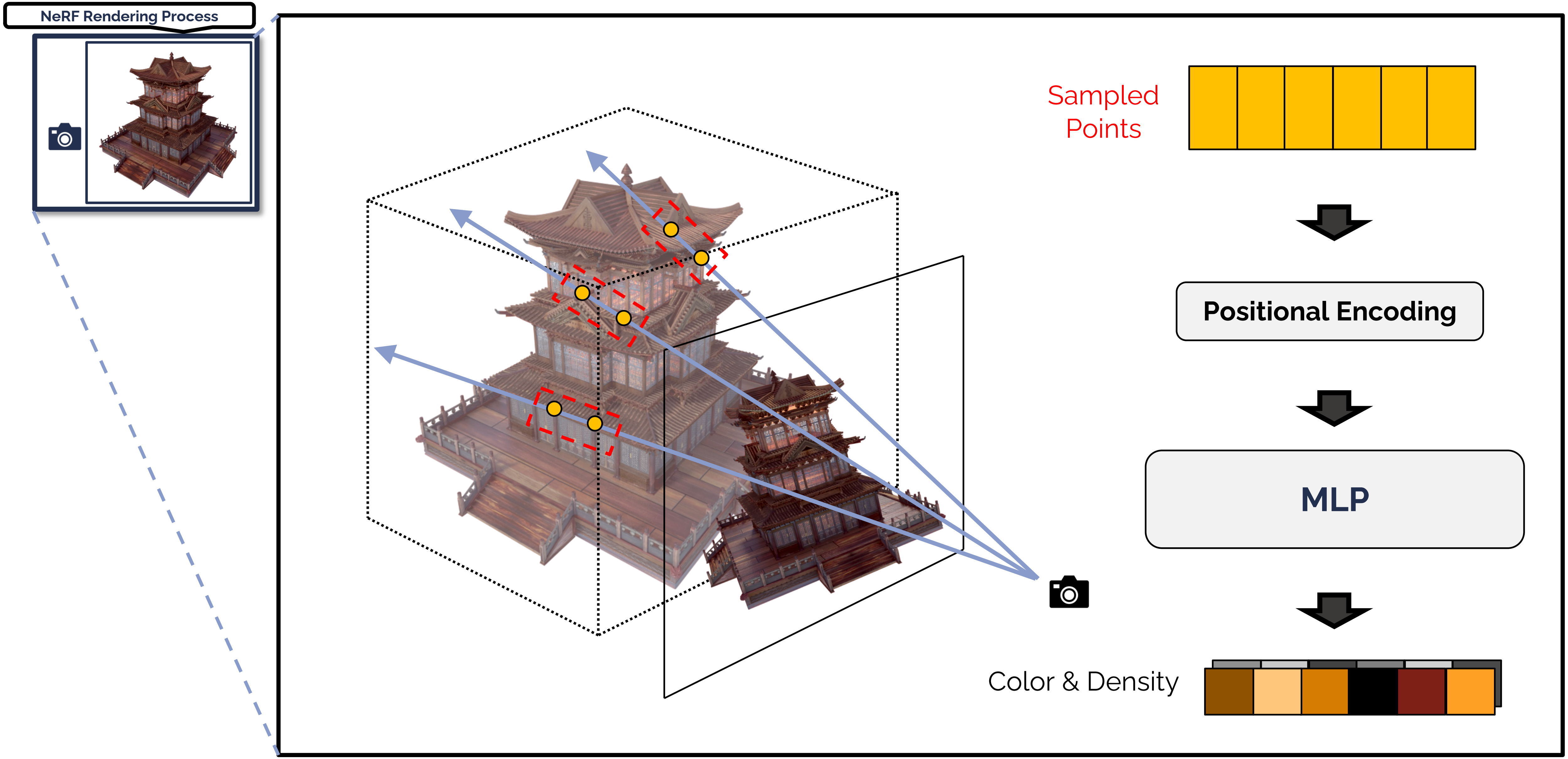




Sampled Points

Positional Encoding

NeRF Rendering Process



NeRF Rendering Process

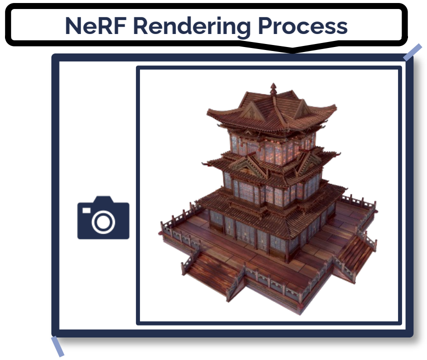
Sampled Points

Positional Encoding

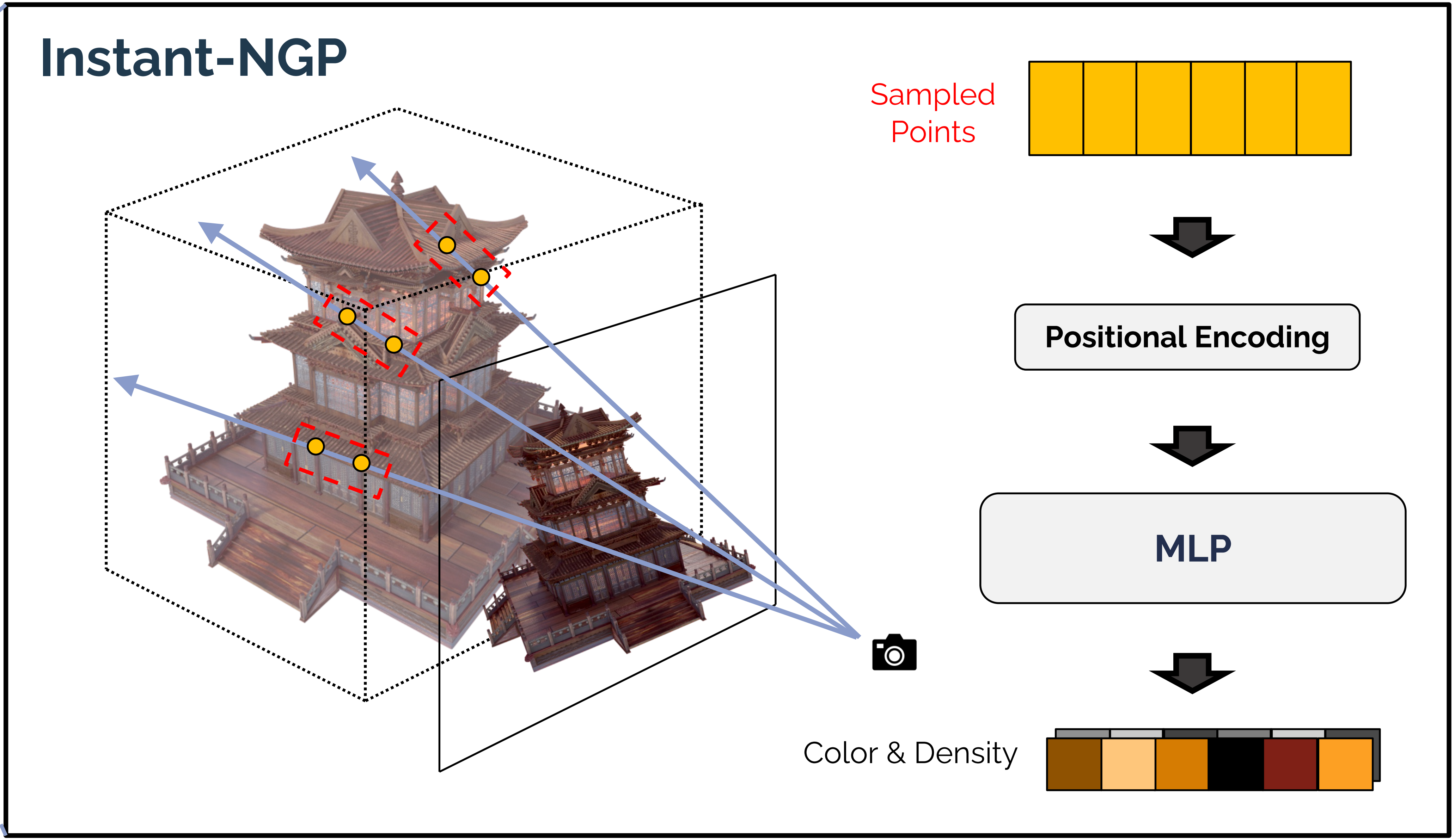
MLP

Color & Density

NeRF Rendering Process



NeRF Rendering Process



Instant-NGP

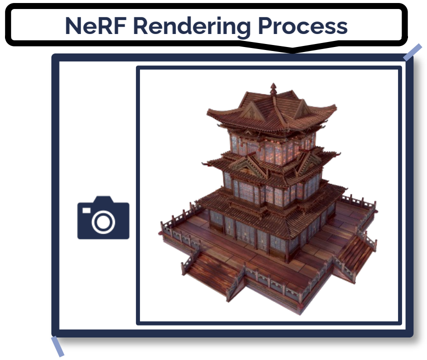
Sampled Points

Positional Encoding

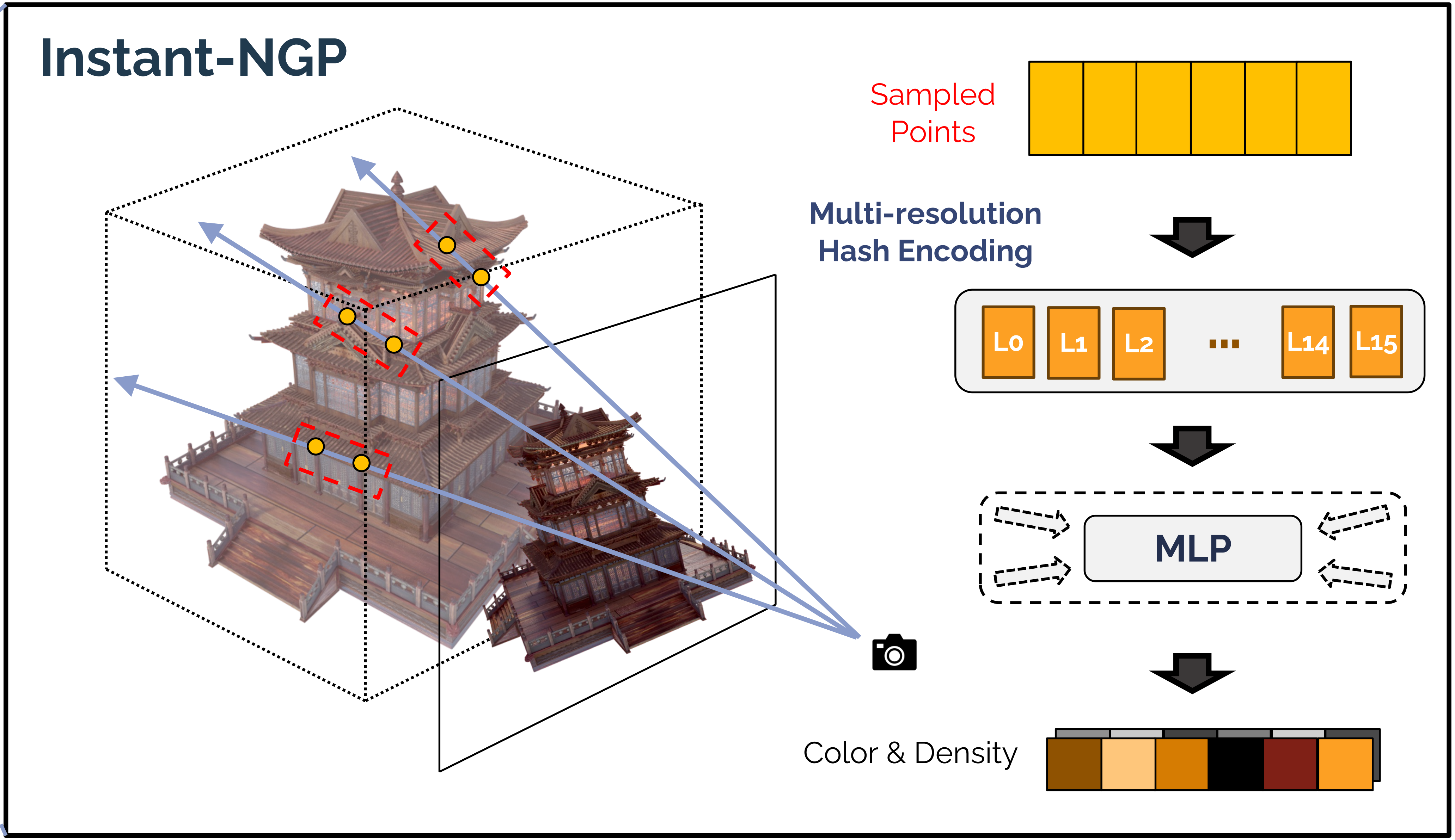
MLP

Color & Density

NeRF Rendering Process



NeRF Rendering Process



Instant-NGP

Sampled Points

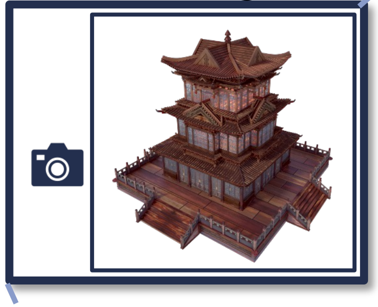
Multi-resolution Hash Encoding

MLP

Color & Density

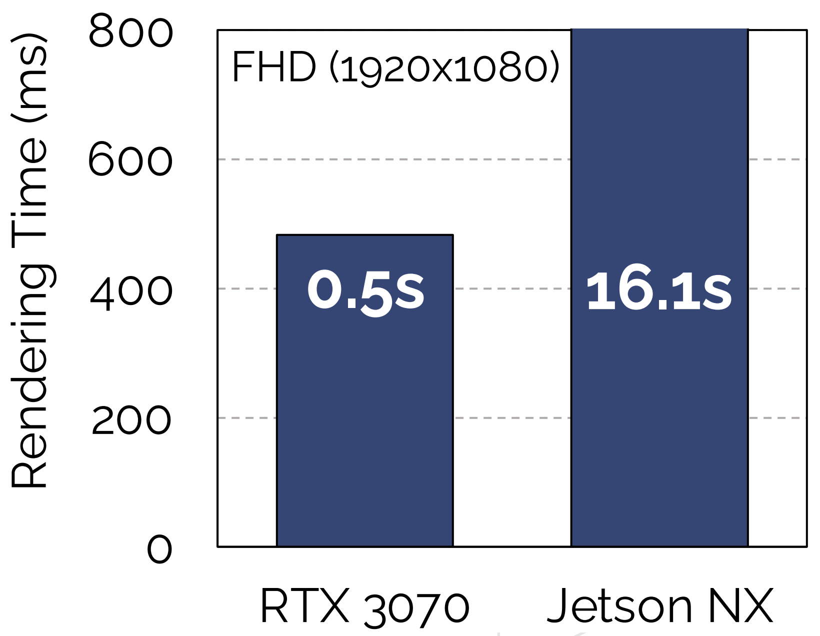
NeRF Rendering Process

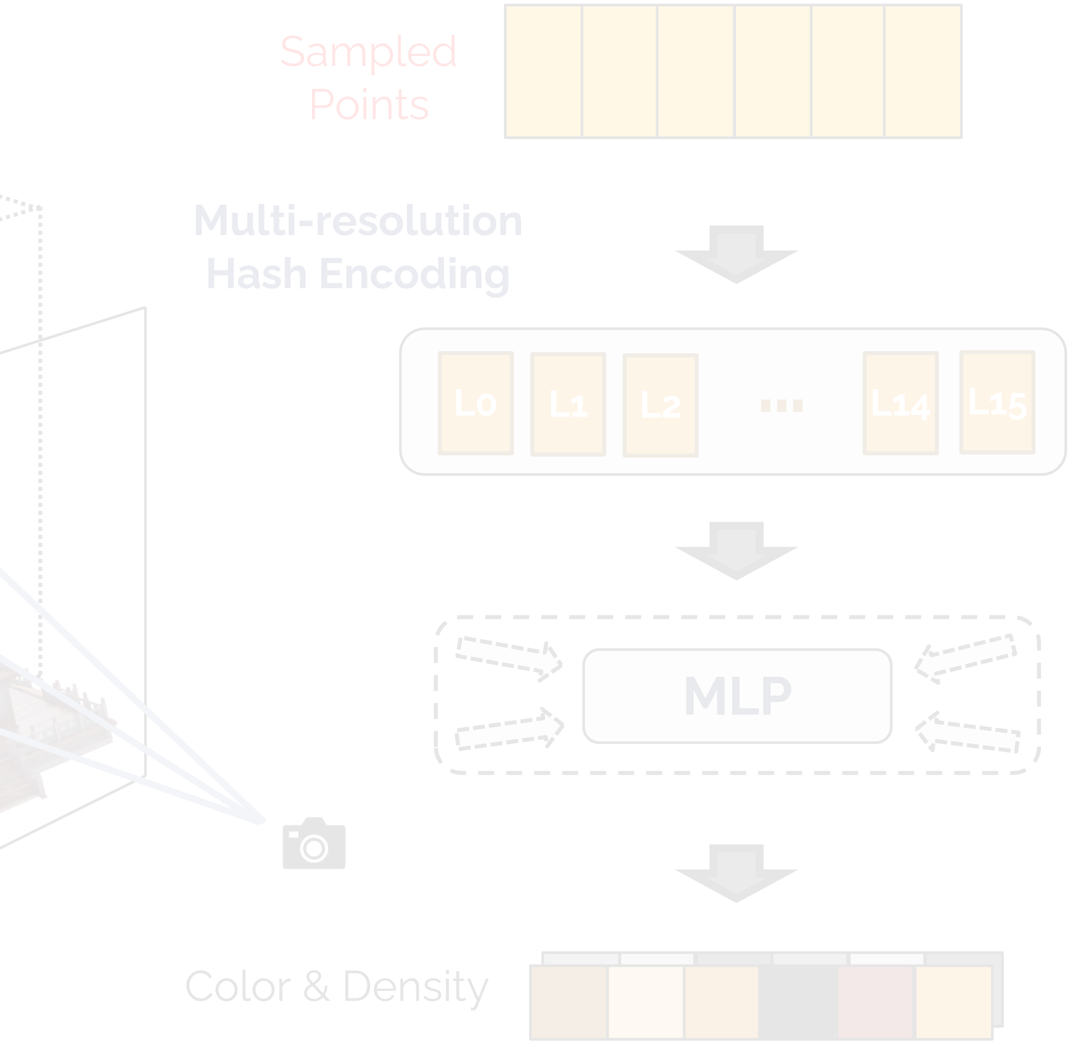
NeRF Rendering Process



Instant-NGP

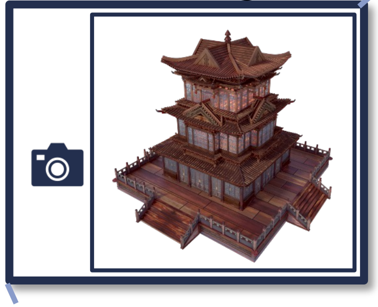
1) Still **too slow**! ☹





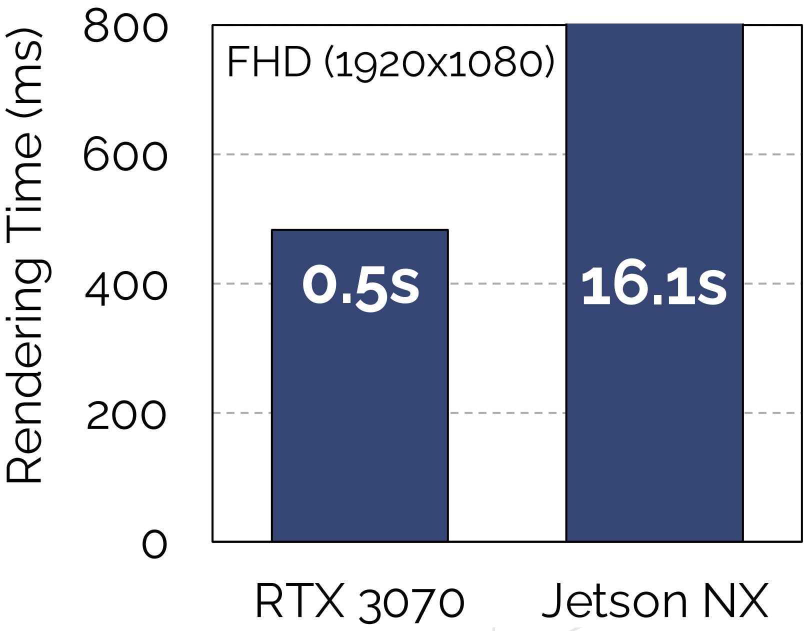
NeRF Rendering Process

NeRF Rendering Process

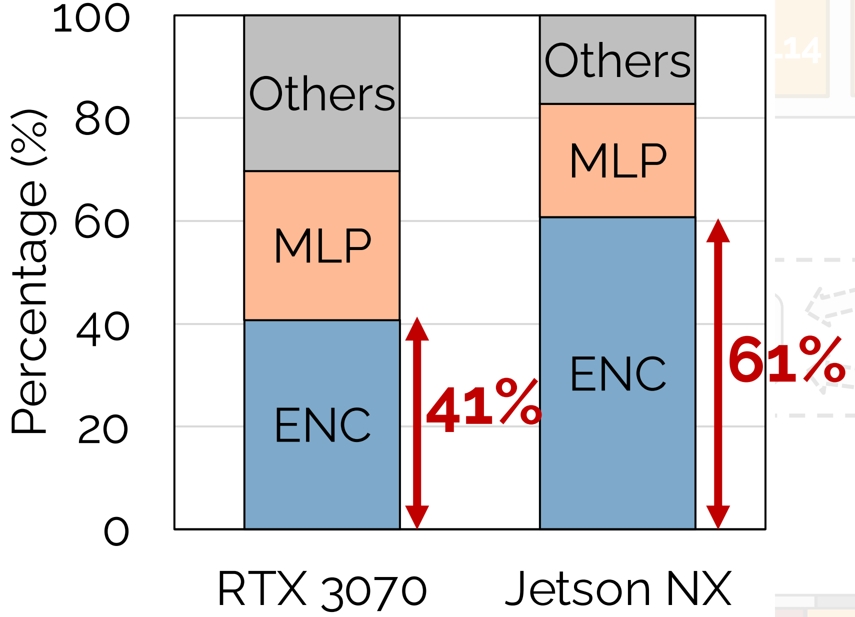


Instant-NGP

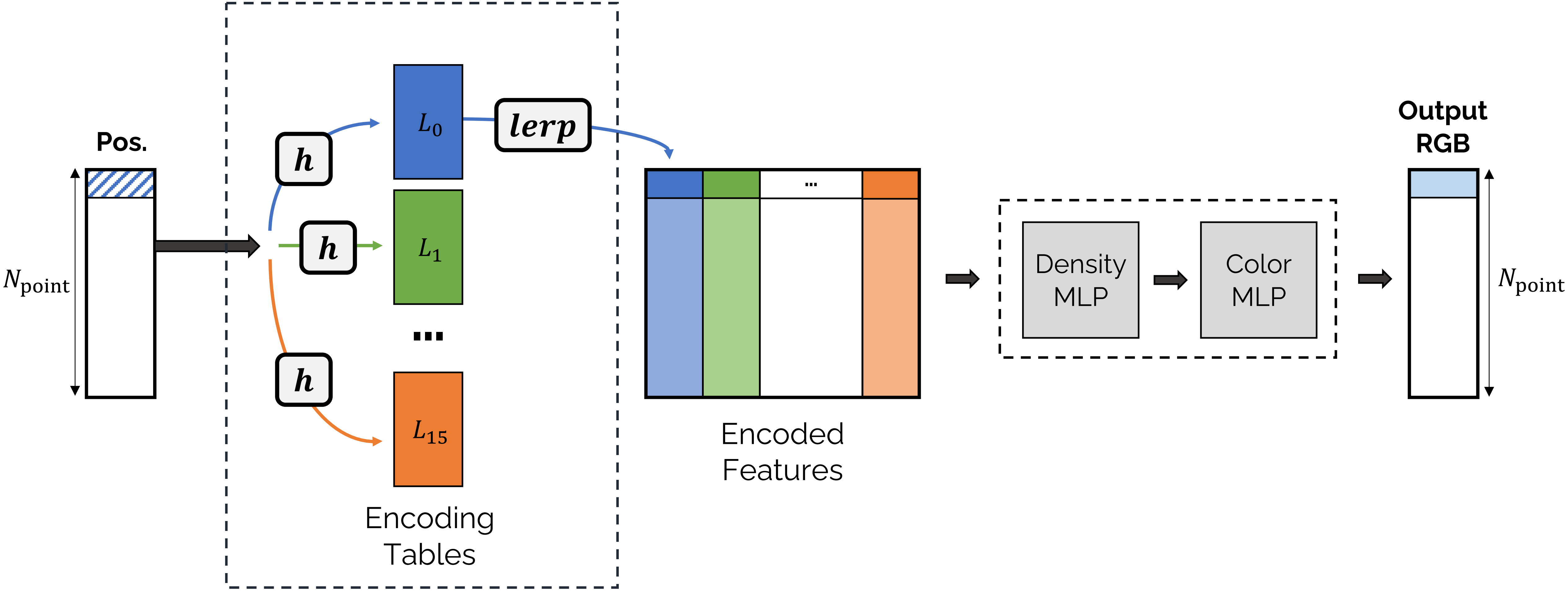
1) Still **too slow**! ☹



2) **Encoding lookup** is bottleneck!

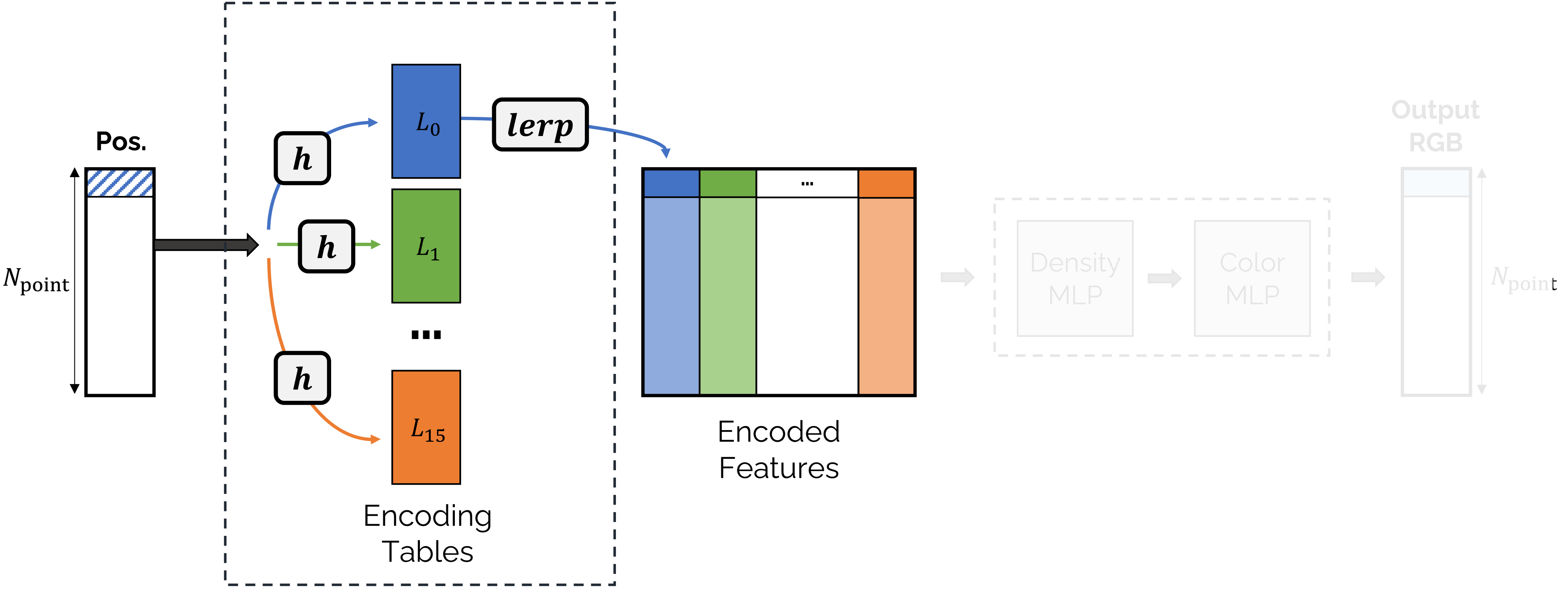


Execution Flow

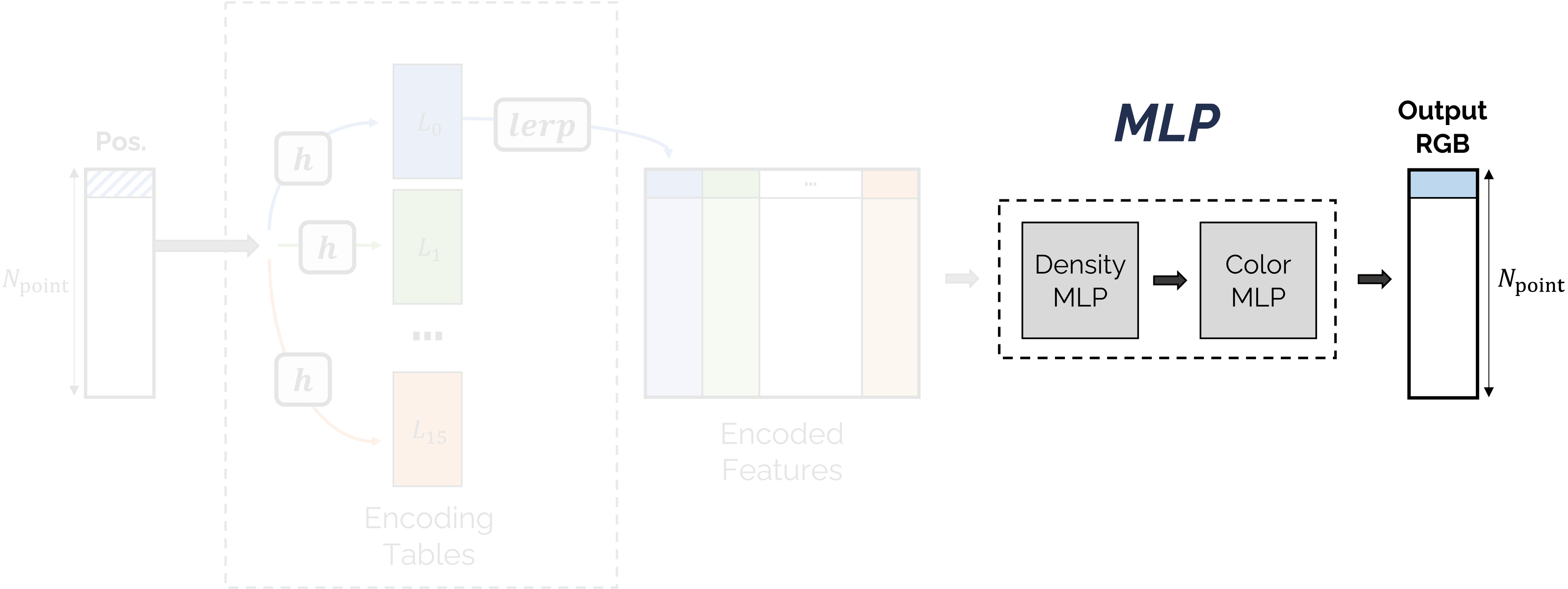


Execution Flow

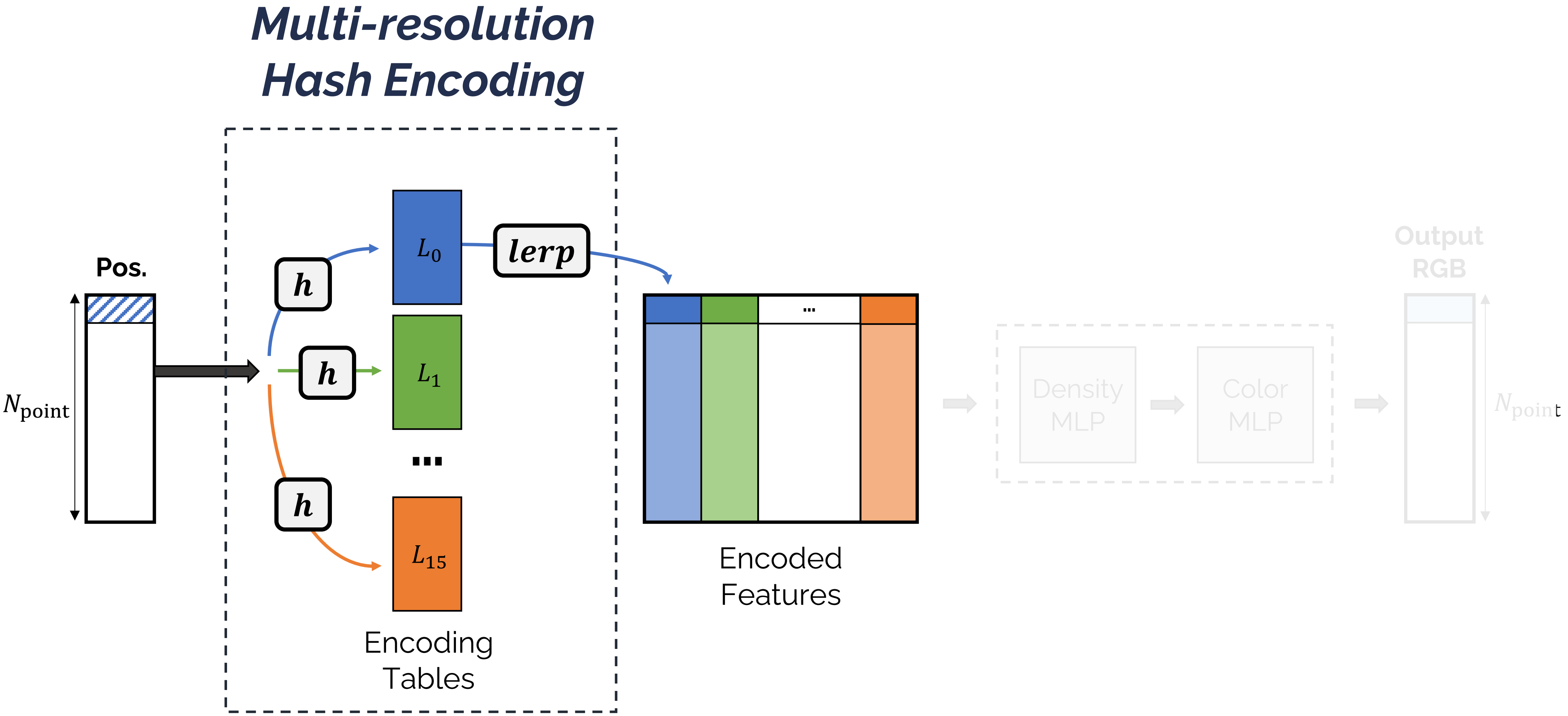
Multi-resolution Hash Encoding



Execution Flow

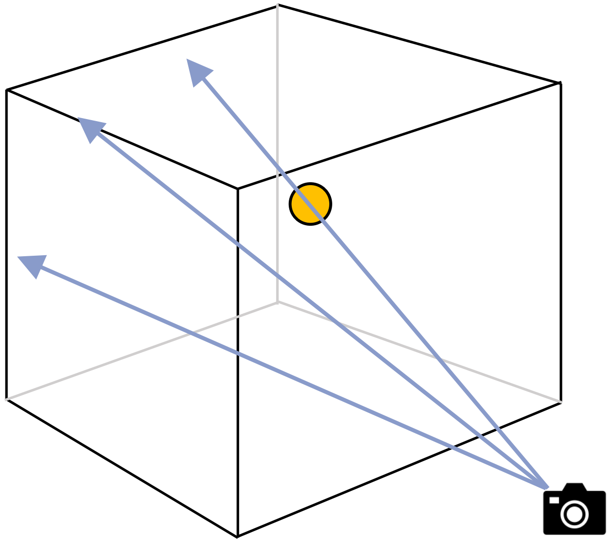


Execution Flow



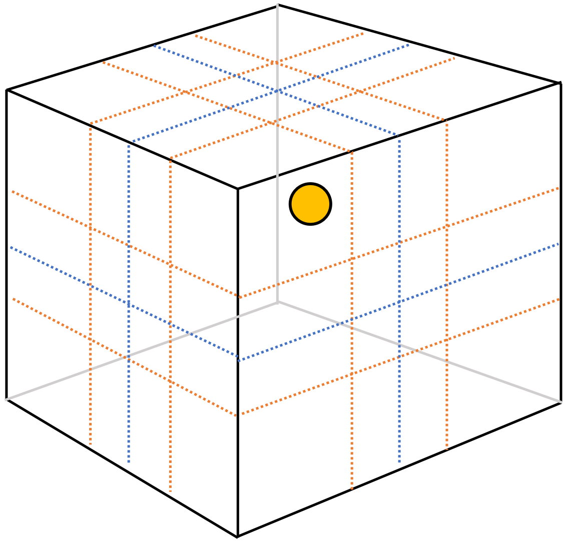
Multi-resolution Hash Encoding

3D Scene (Bounding Box)



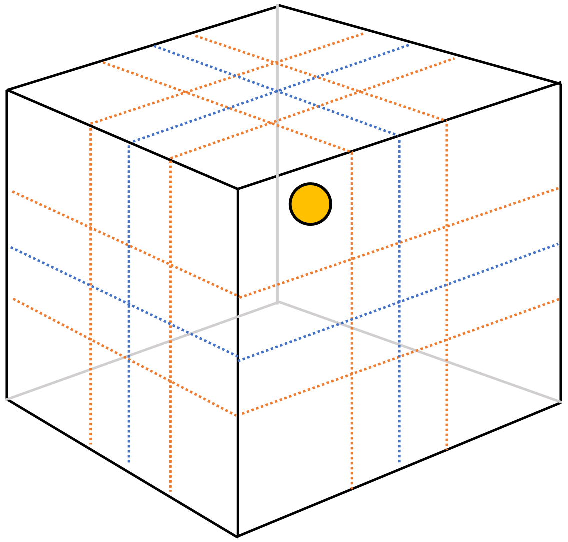
Multi-resolution Hash Encoding

3D Scene (Bounding Box)



Multi-resolution Hash Encoding

3D Scene (Bounding Box)

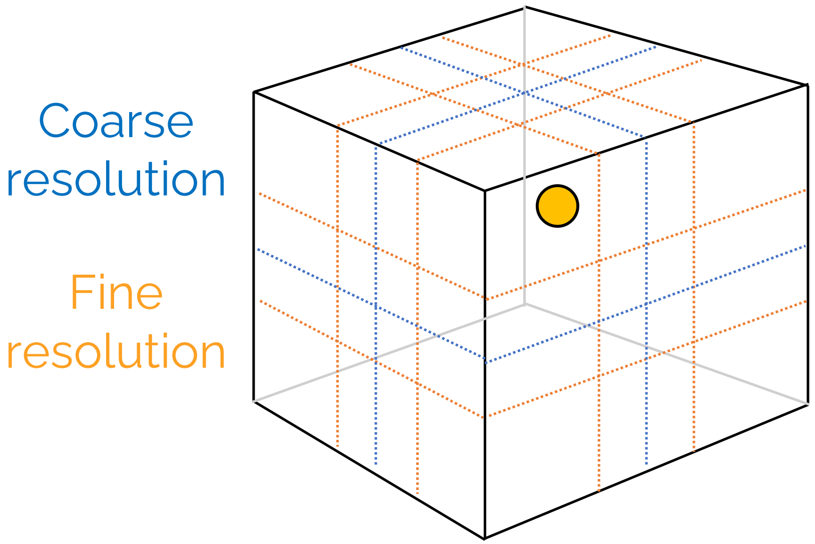


Coarse resolution

Fine resolution

Multi-resolution Hash Encoding

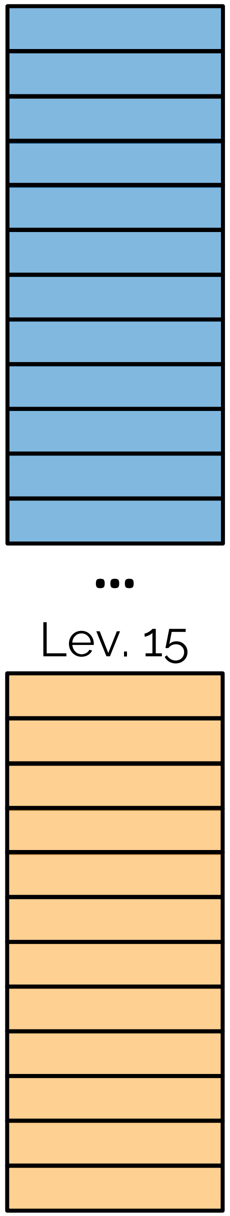
3D Scene (Bounding Box)



Coarse resolution

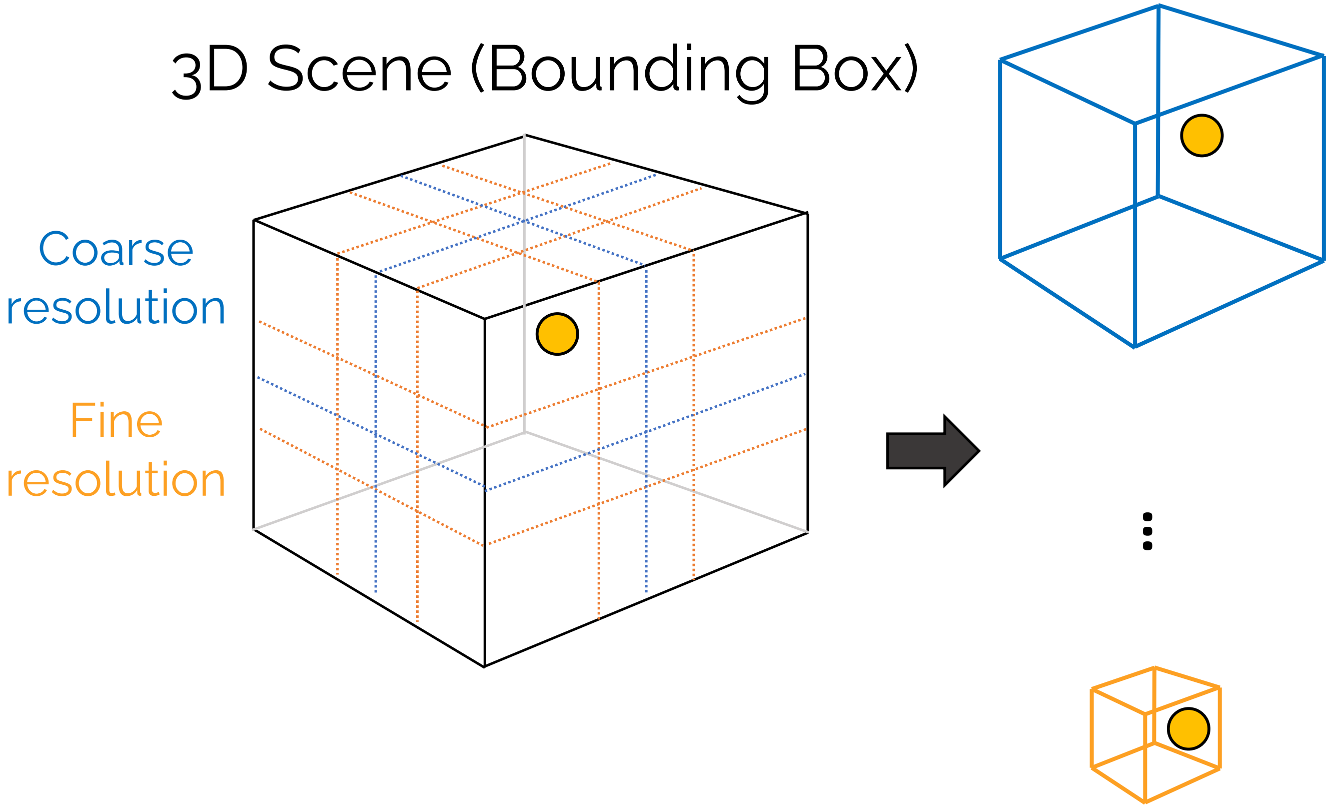
Fine resolution

Lev. 0

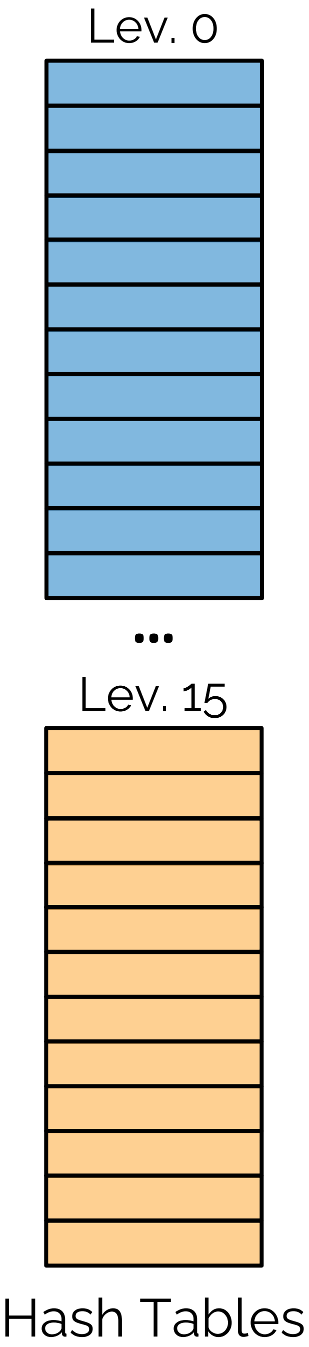


Hash Tables

Multi-resolution Hash Encoding

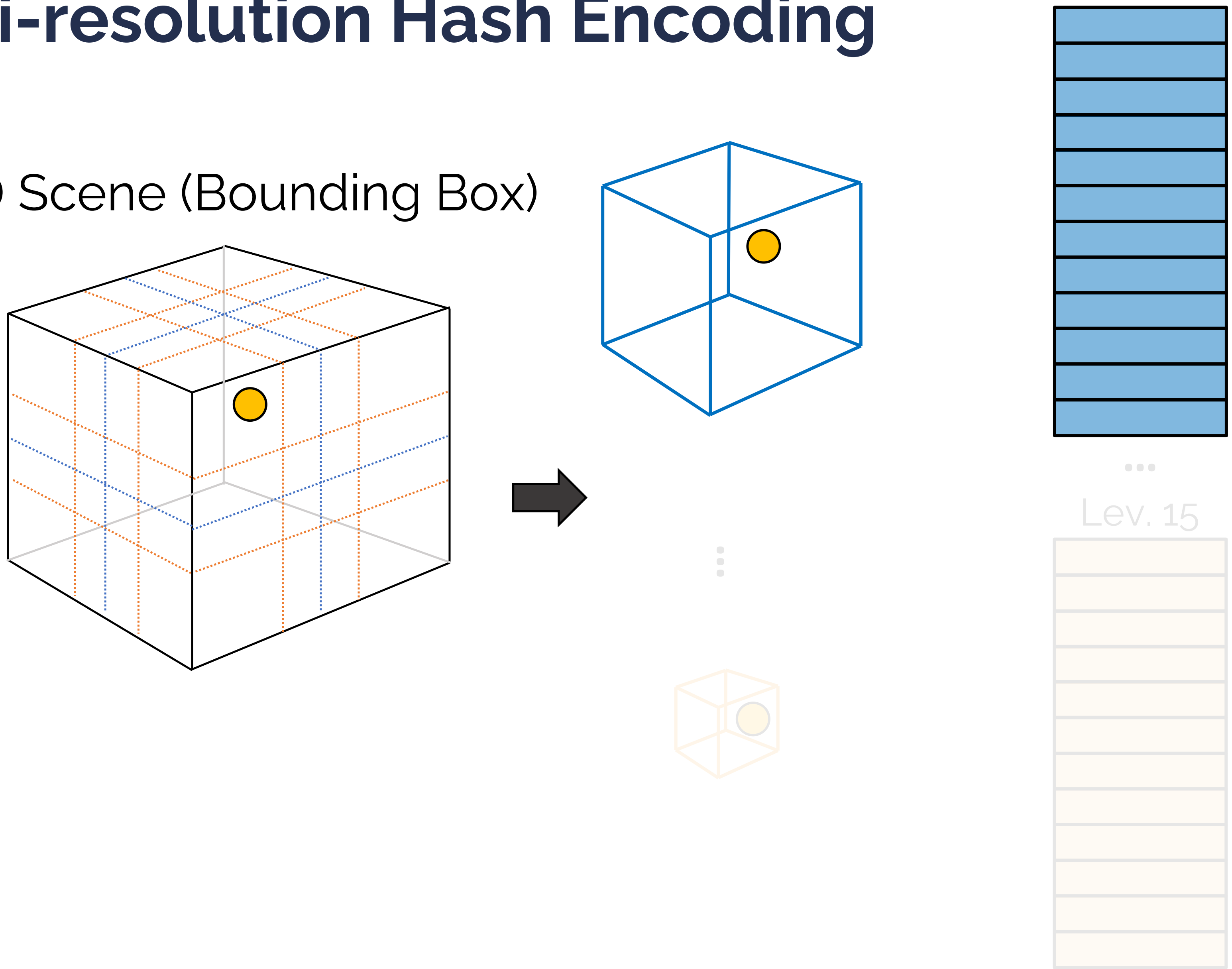


3D Scene (Bounding Box)



Hash Tables

Multi-resolution Hash Encoding



3D Scene (Bounding Box)

Coarse resolution

Fine resolution

Lev. 0

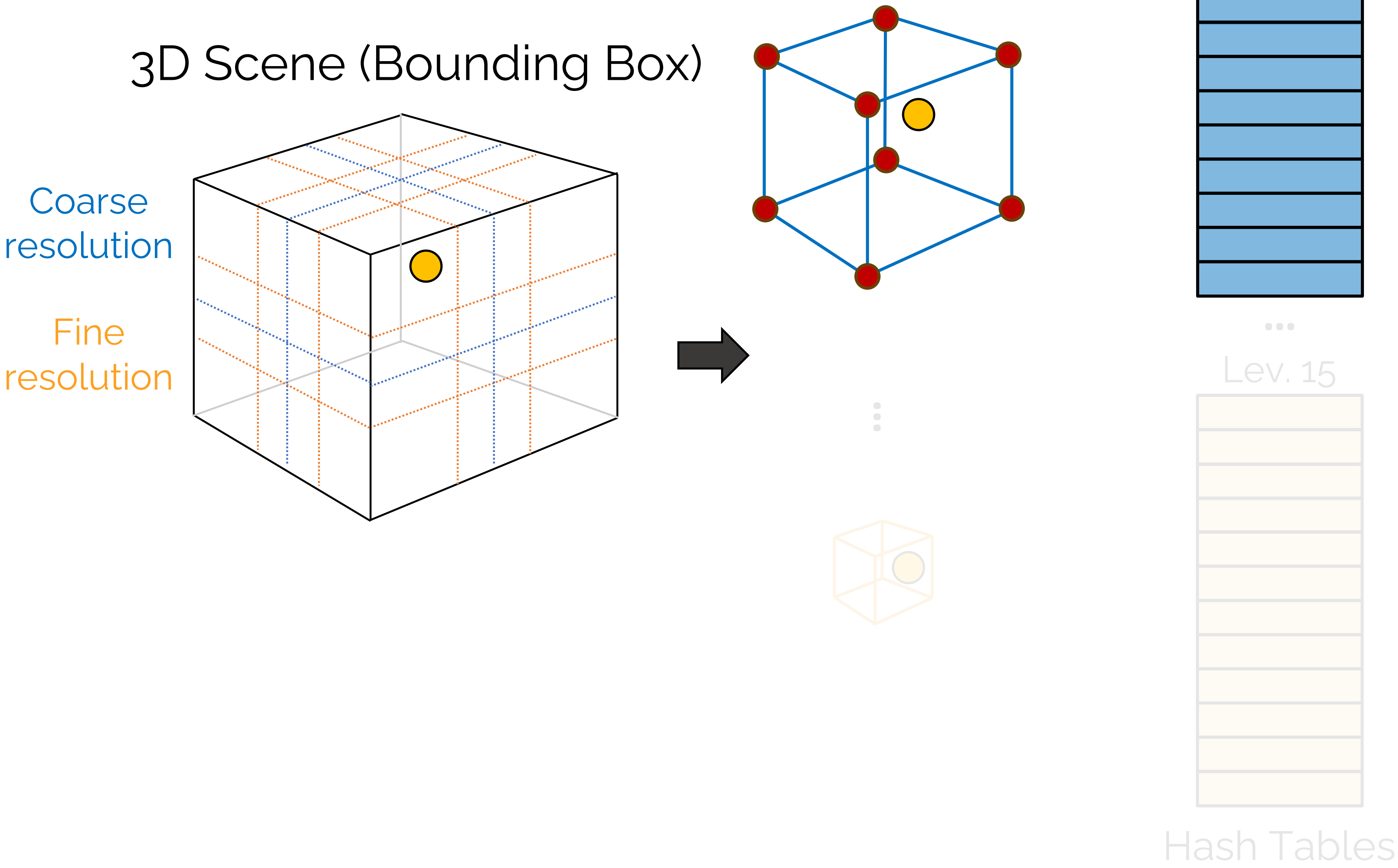
Lev. 15

Hash Tables

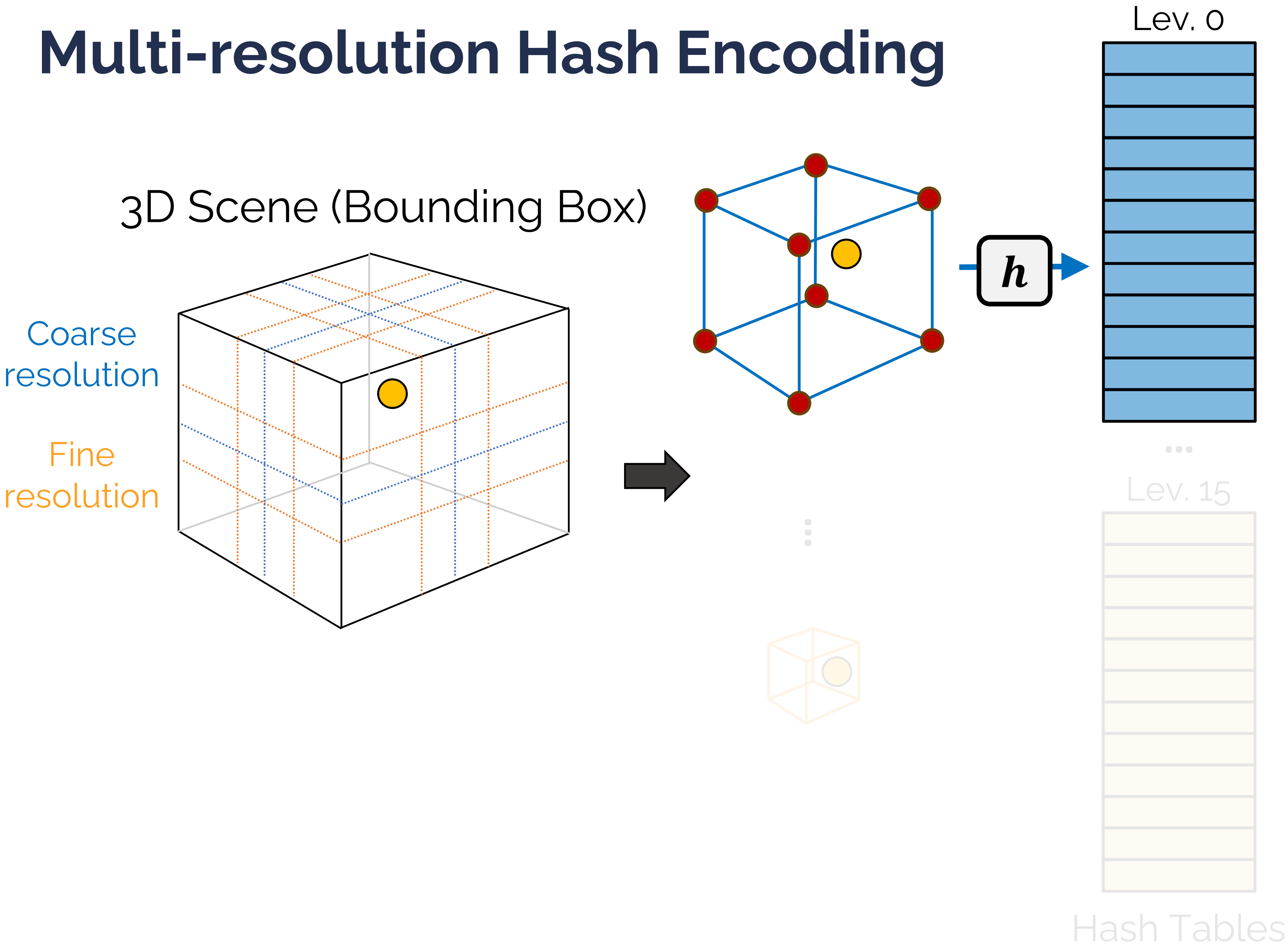
Multi-resolution Hash Encoding

Lev. 0

3D Scene (Bounding Box)



Multi-resolution Hash Encoding



3D Scene (Bounding Box)

Coarse resolution

Fine resolution

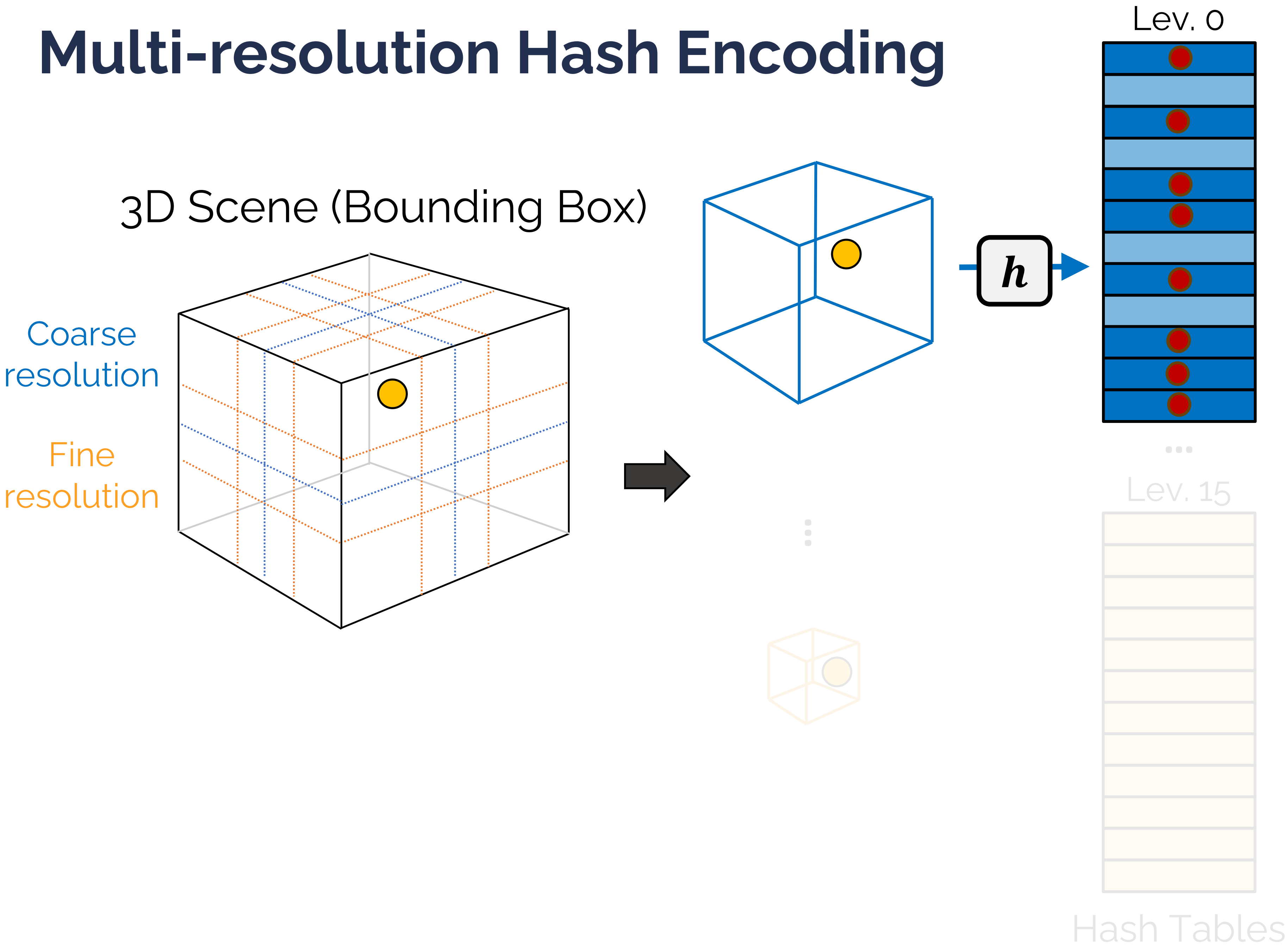
Lev. 0

...

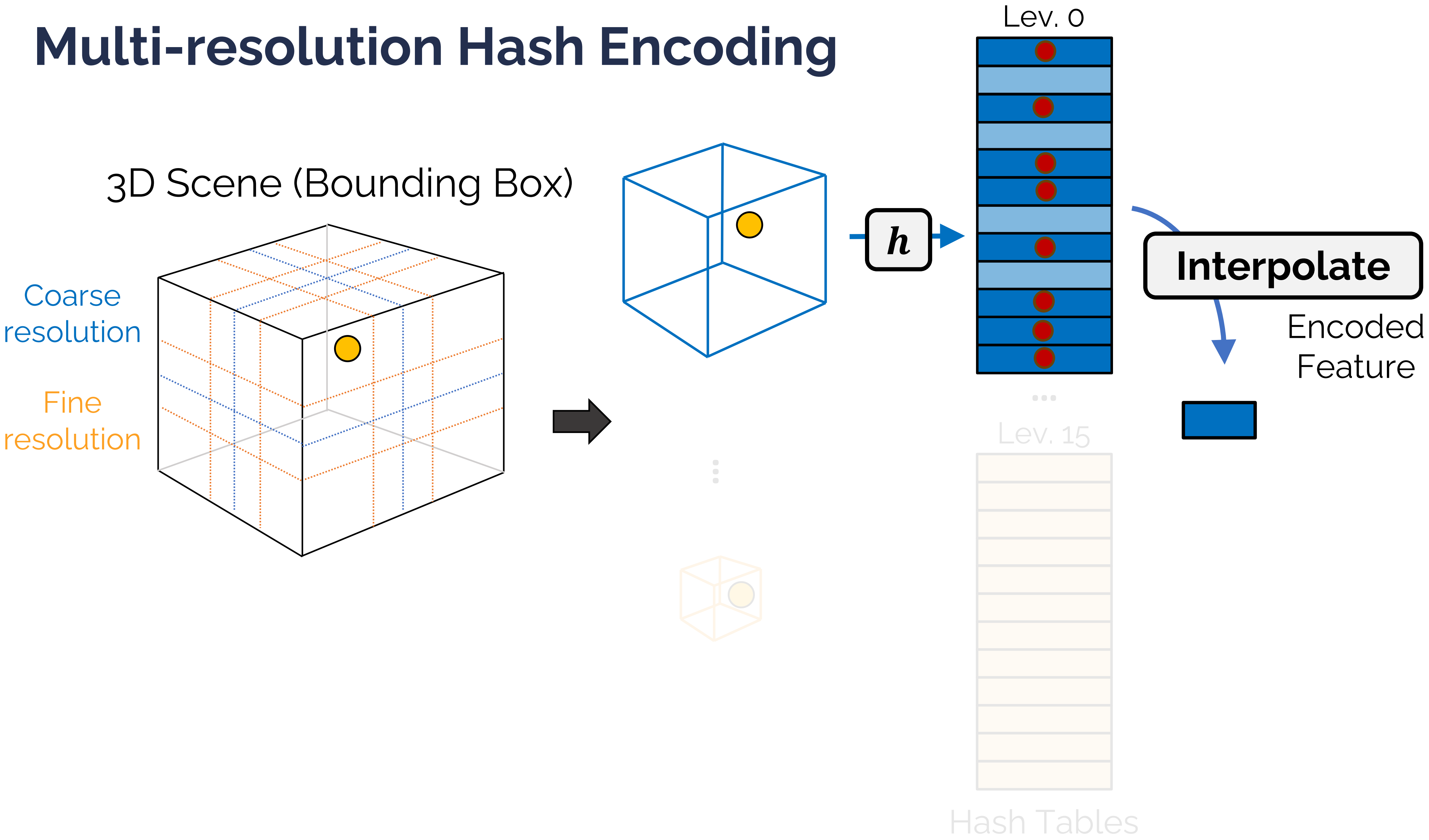
Lev. 15

Hash Tables

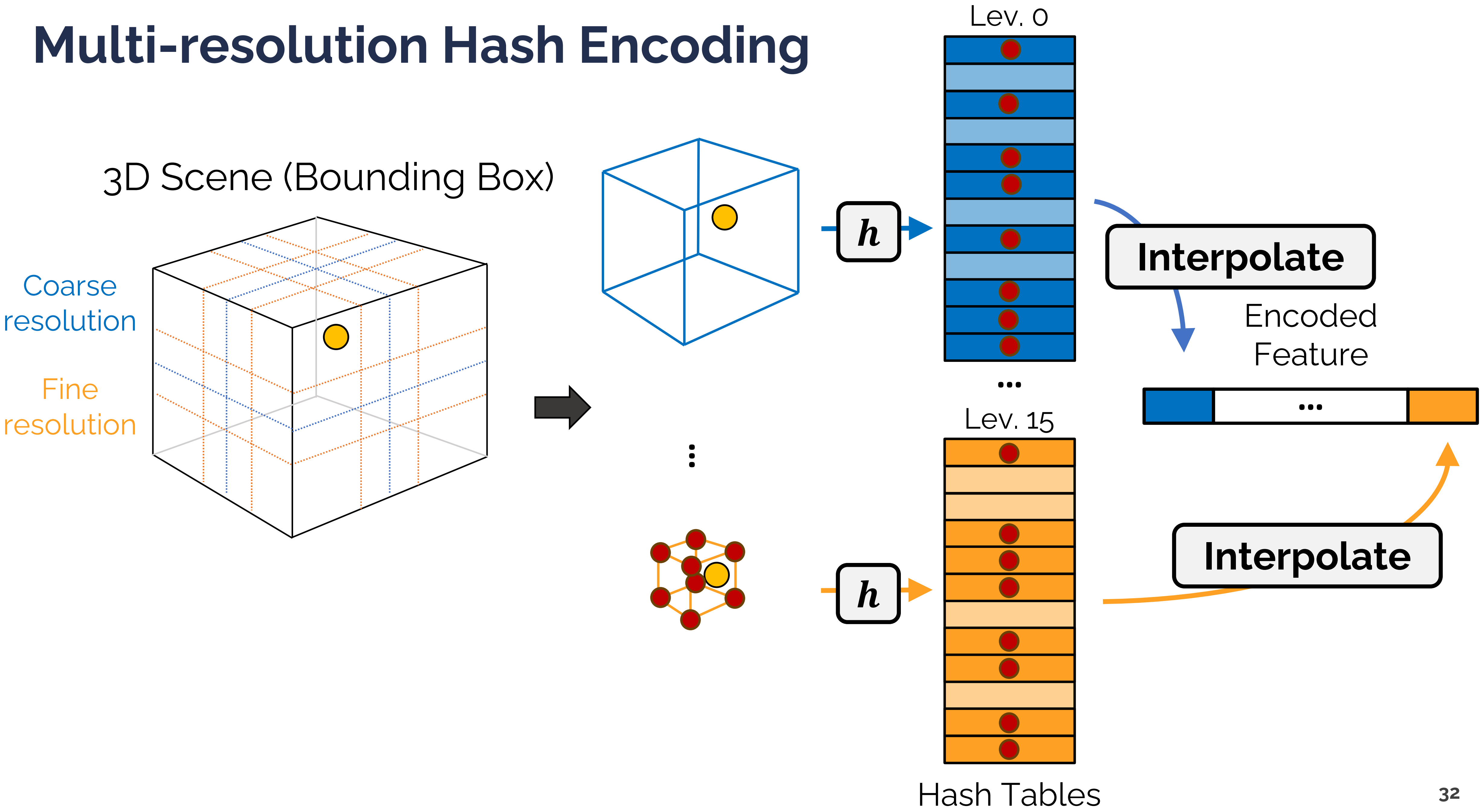
Multi-resolution Hash Encoding



Multi-resolution Hash Encoding



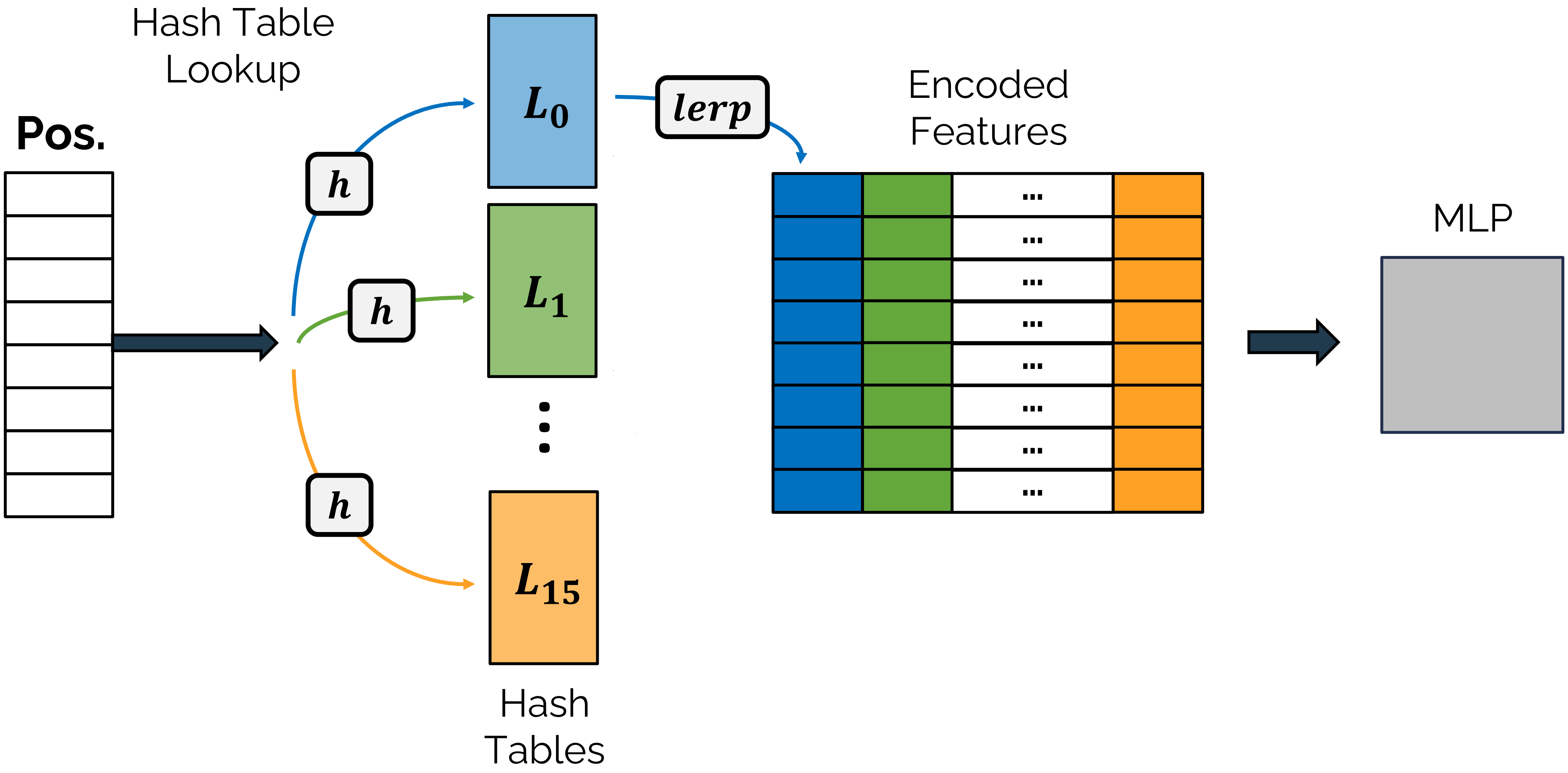
Multi-resolution Hash Encoding



Outline

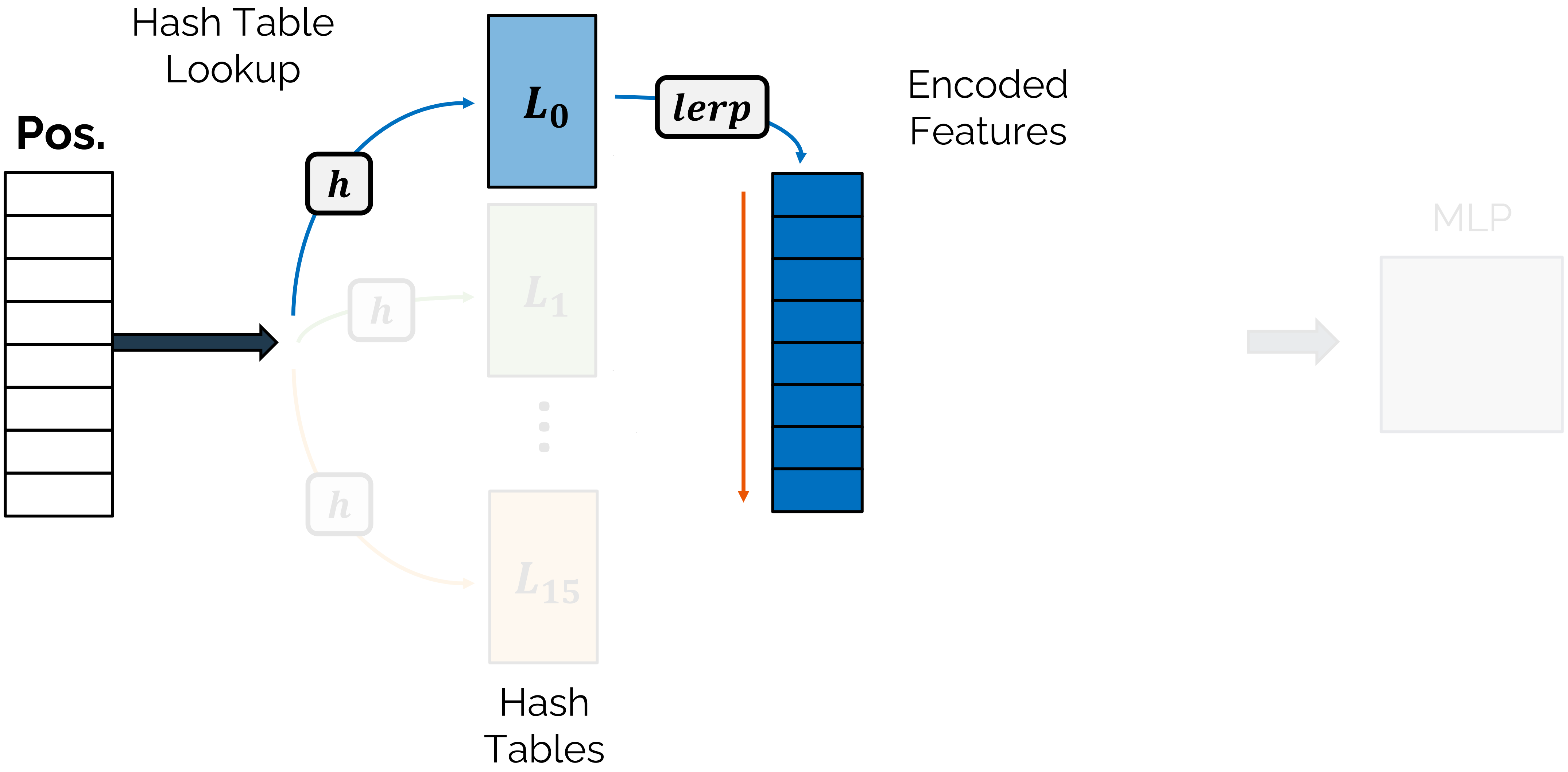
- **Background**
  - Neural Rendering
  - Multi-resolution Hash Encoding
- **I-NGP Optimization & Limitations**
- **NeuRex: Efficient Neural Rendering Accelerator**
  - Restricted Hashing
  - Neural Graphics Engine
- **Evaluation**
- **Conclusion**

I-NGP Optimization



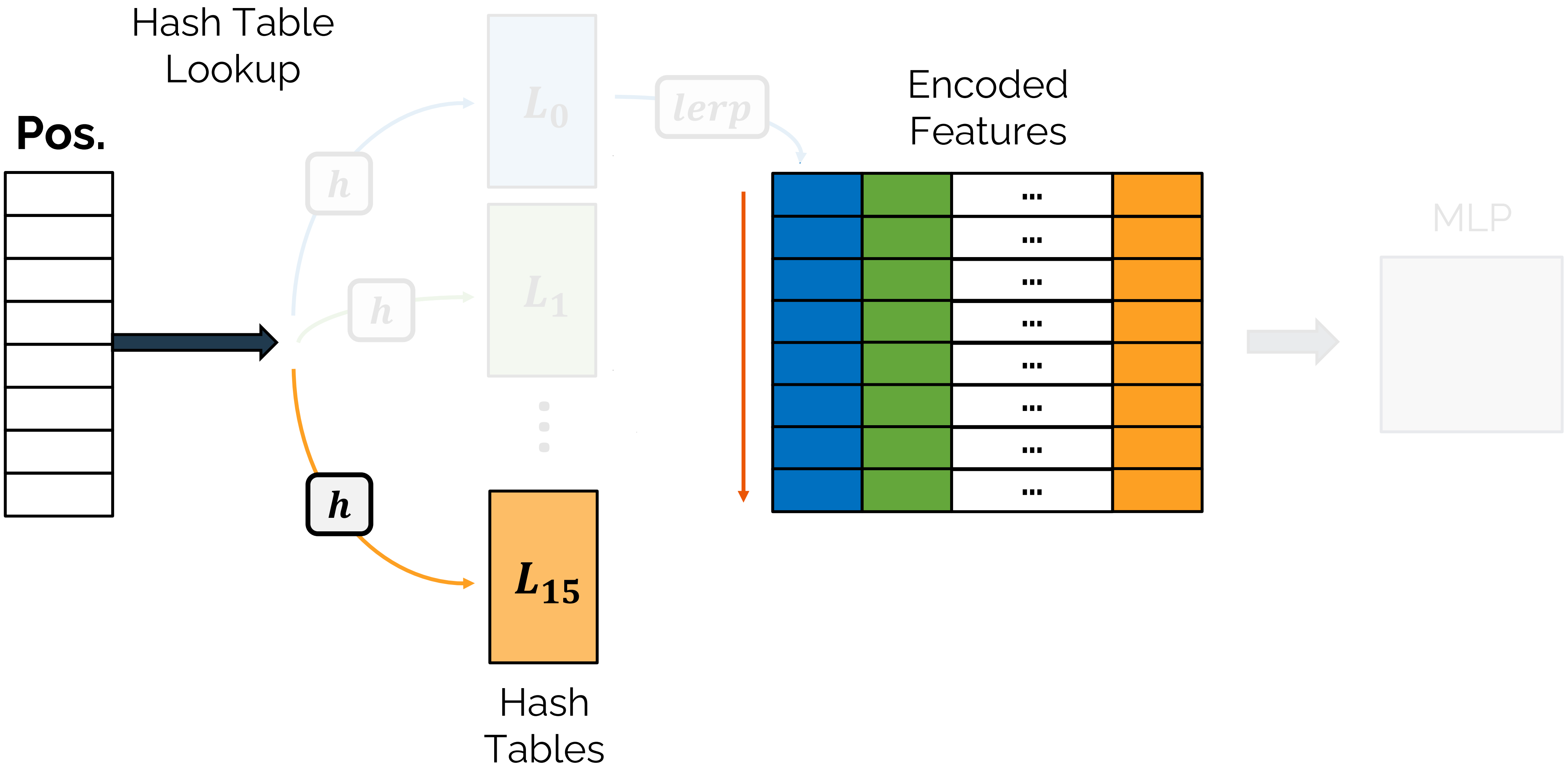
I-NGP Optimization

Level-wise execution



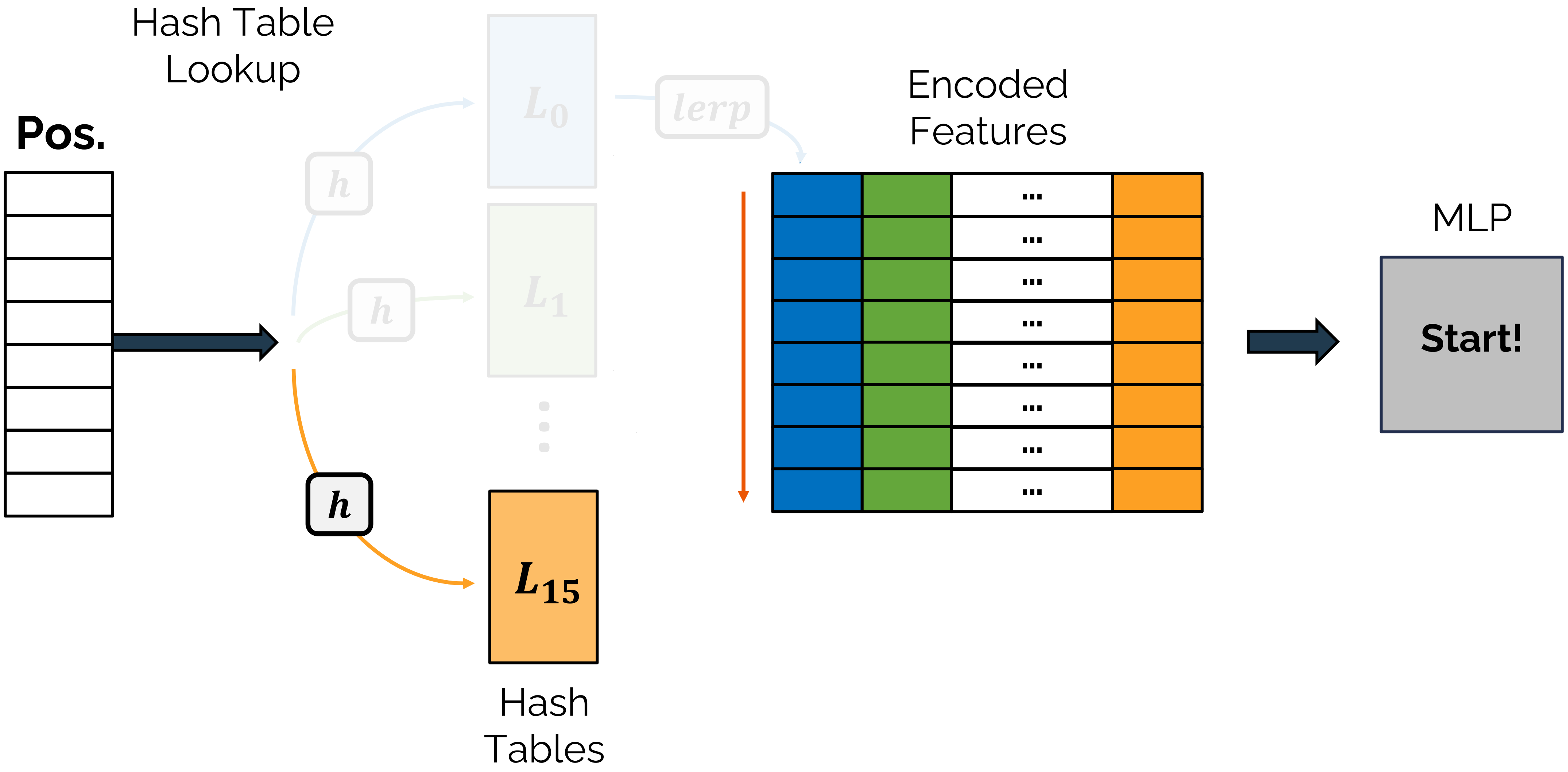
I-NGP Optimization

Level-wise execution

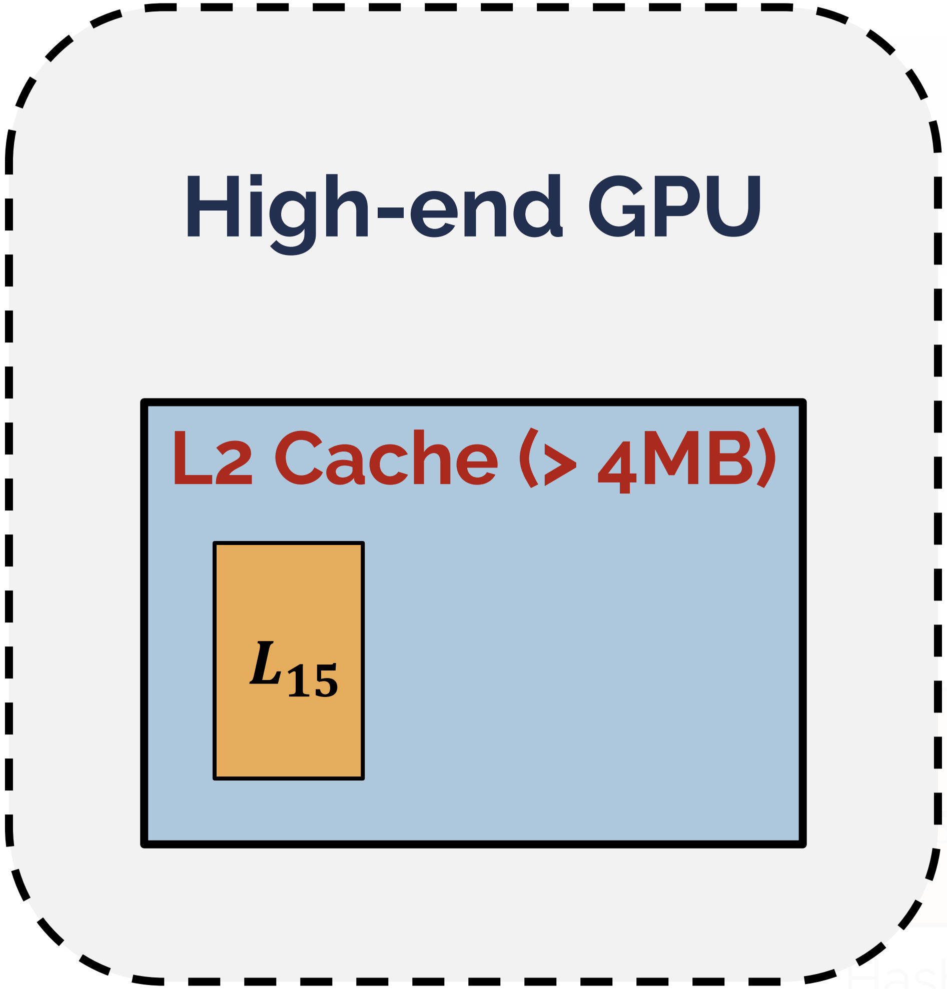


I-NGP Optimization

Level-wise execution



I-NGP Optimization



High-end GPU

L2 Cache (> 4MB)

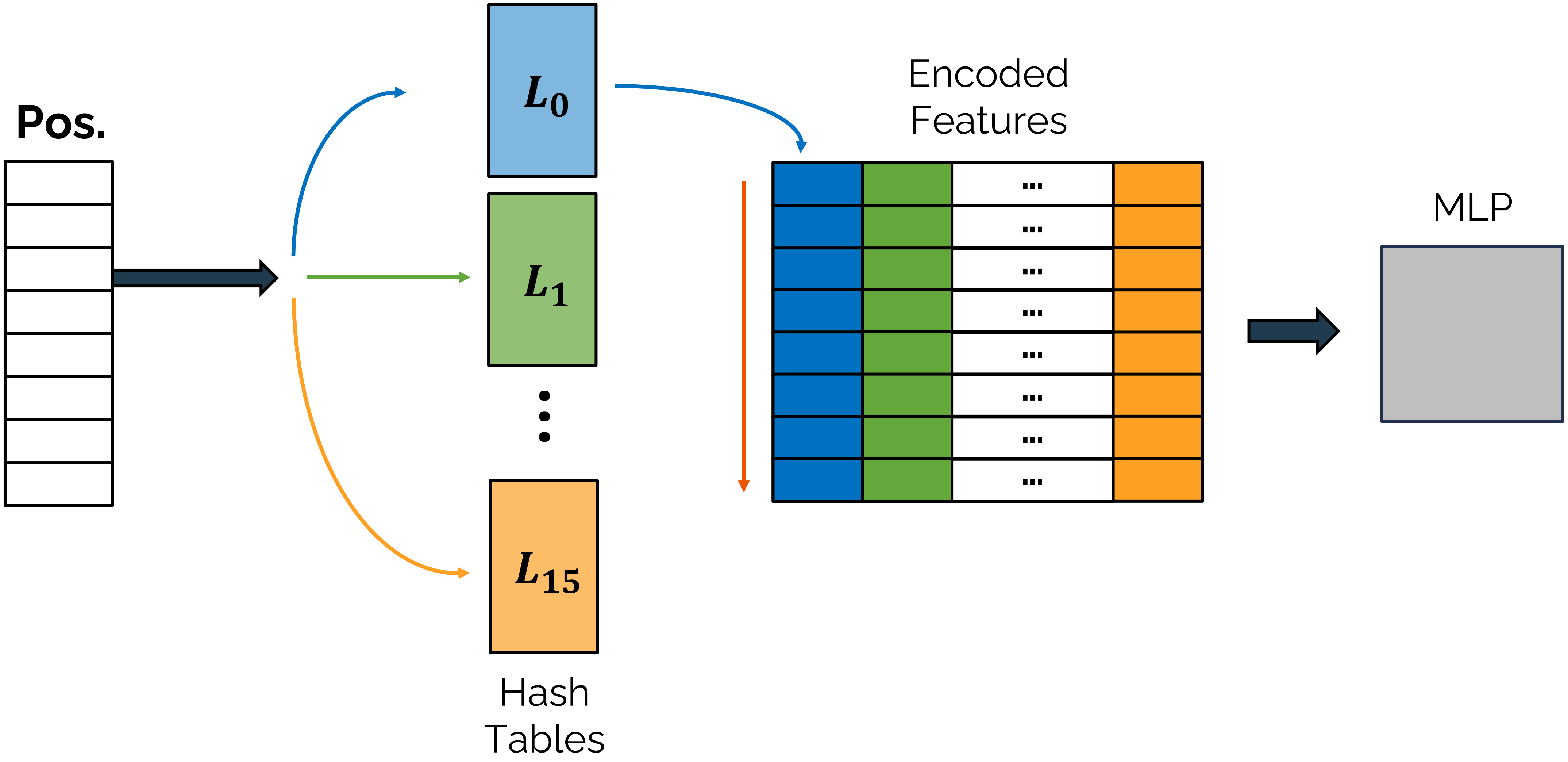
L_{15}

- L2 cache in high-end GPUs > Single hash table

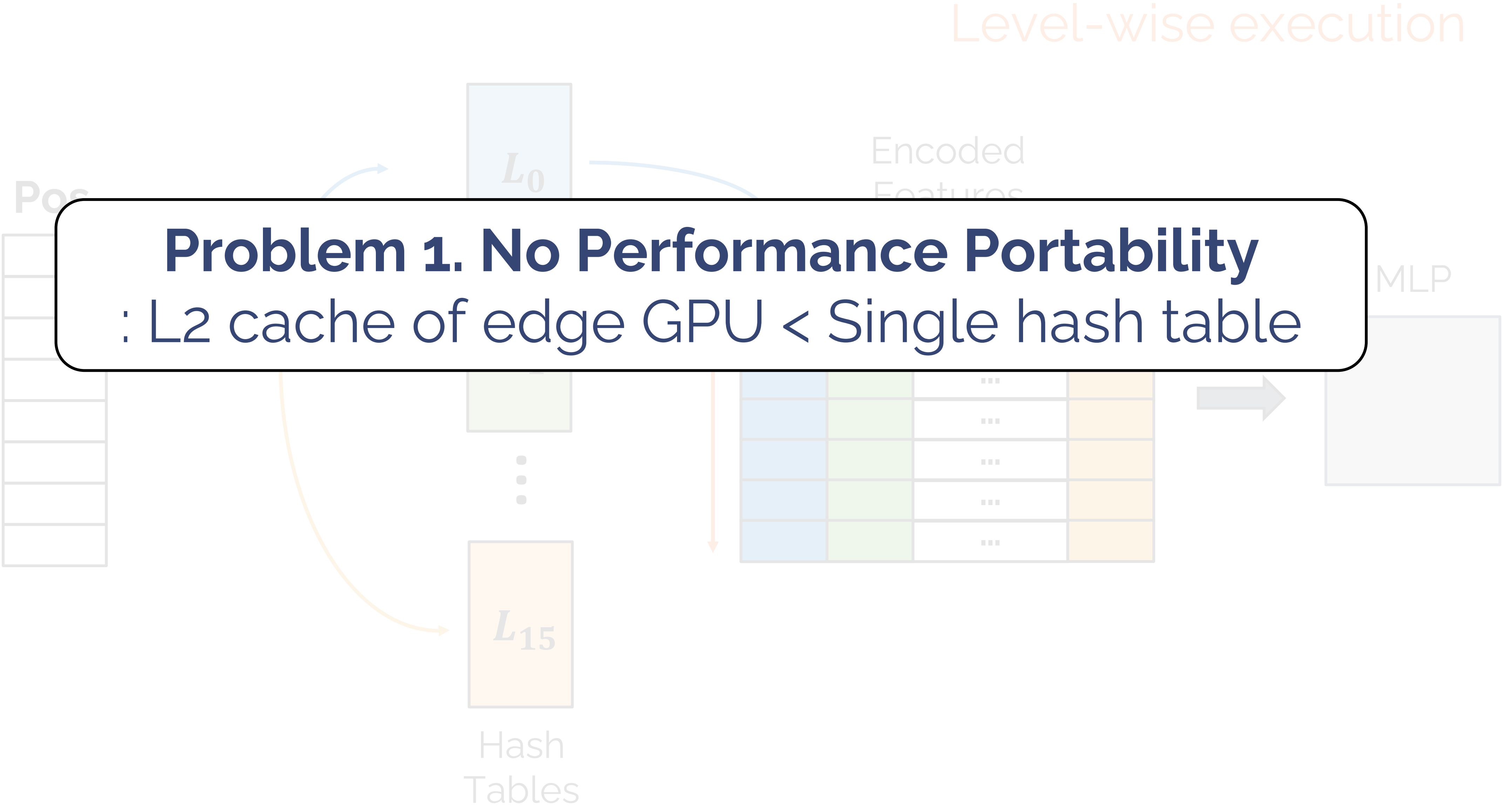
- Implication
  - Minimize off-chip memory access

Limitations of I-NGP Optimization

Level-wise execution



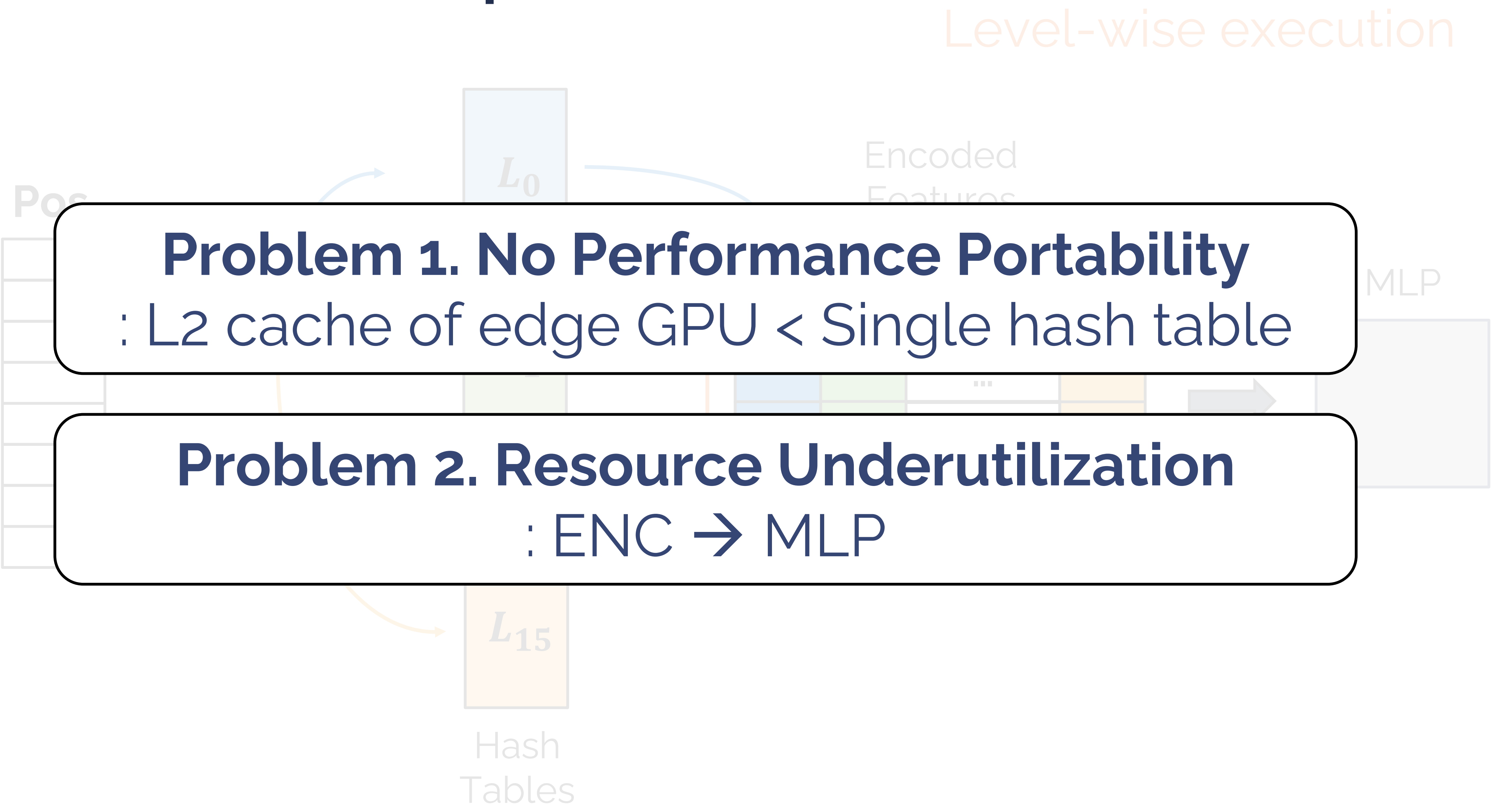
Limitations of I-NGP Optimization



Problem 1. No Performance Portability

: L2 cache of edge GPU < Single hash table

Limitations of I-NGP Optimization



Problem 1. No Performance Portability

: L2 cache of edge GPU < Single hash table

Problem 2. Resource Underutilization

: ENC → MLP

Outline

- **Background**
  - Neural Rendering
  - Multi-resolution Hash Encoding

- **I-NGP Optimization & Limitations**

- **NeuRex: Efficient Neural Rendering Accelerator**
  - Restricted Hashing
  - Neural Graphics Engine

- **Evaluation**

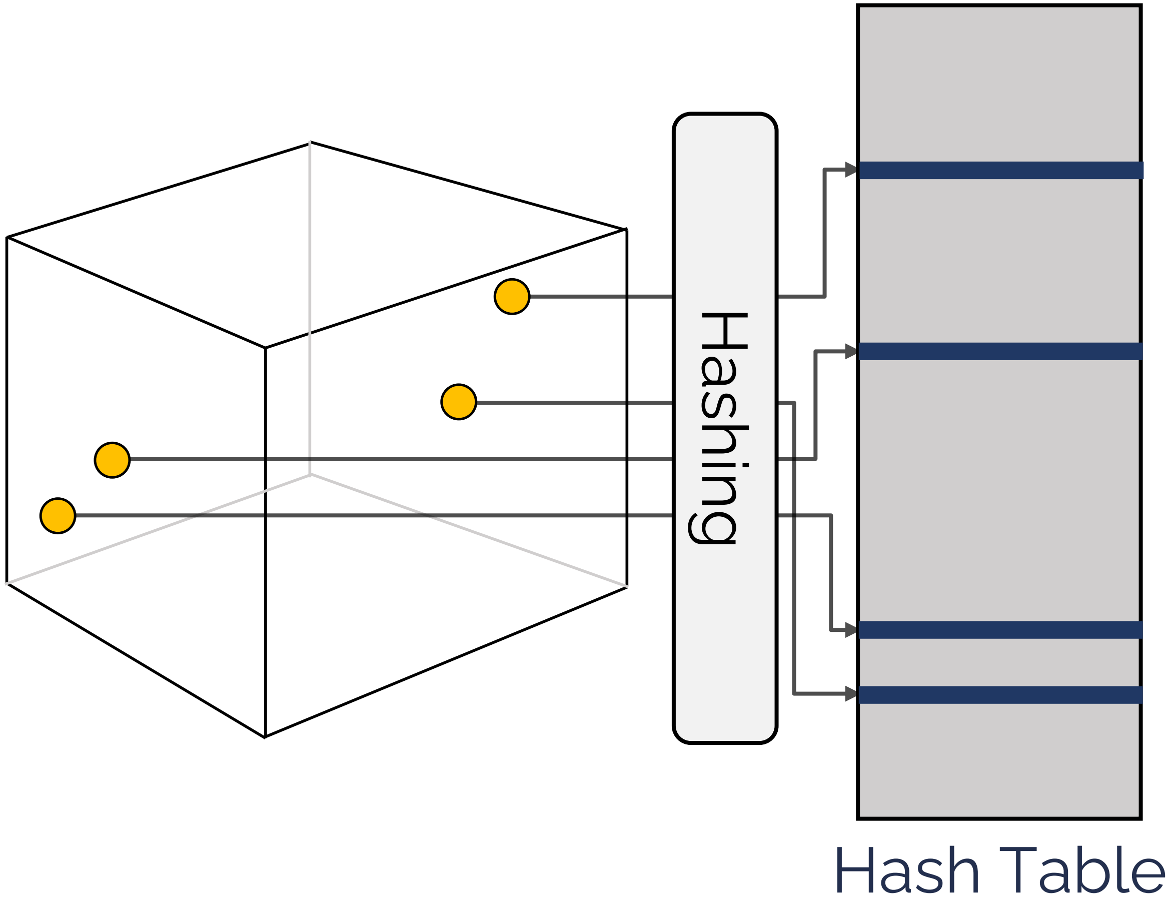
- **Conclusion**

NeuRex: Hardware-Friendly Hash Encoding

NeuRex: Hardware-Friendly Hash Encoding

Original Algorithm

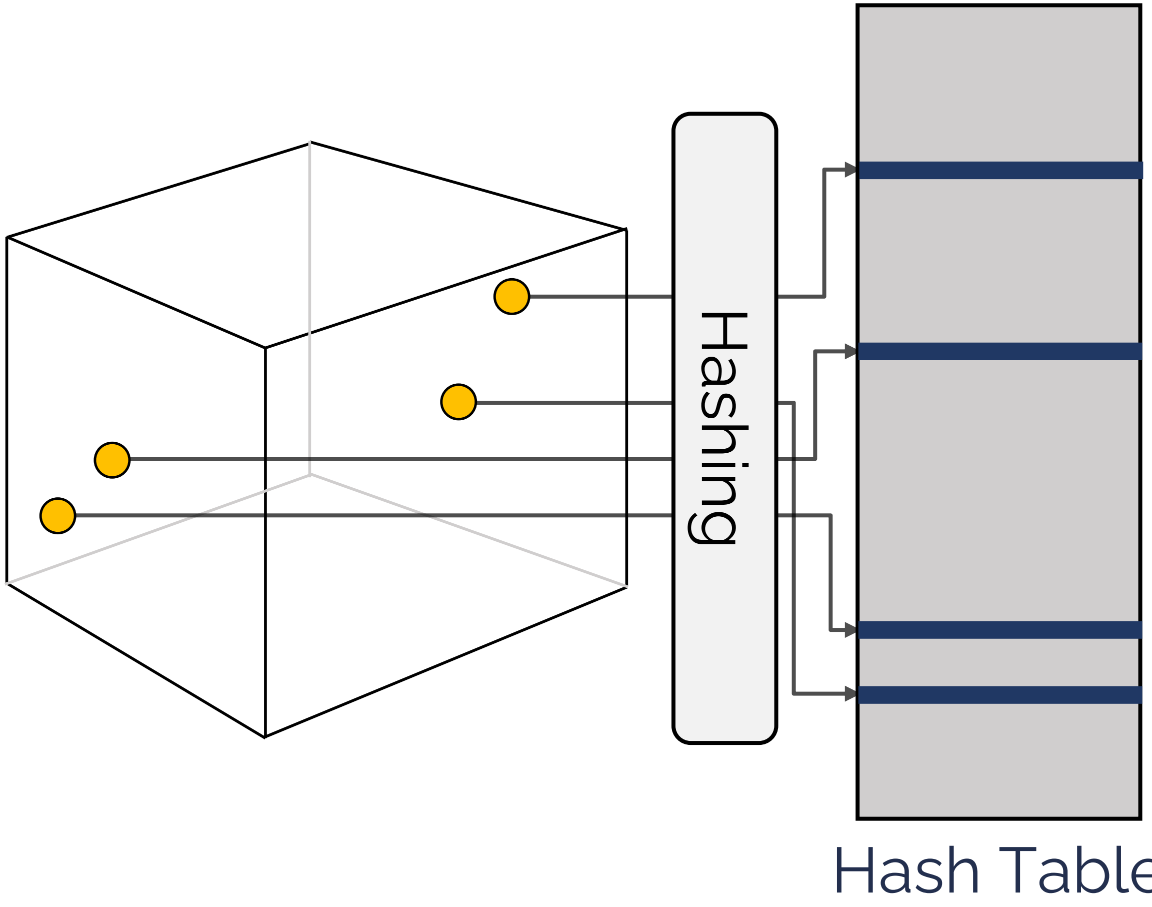
- Access range: **Entire hash table ☹**



NeuRex: Hardware-Friendly Hash Encoding

Original Algorithm

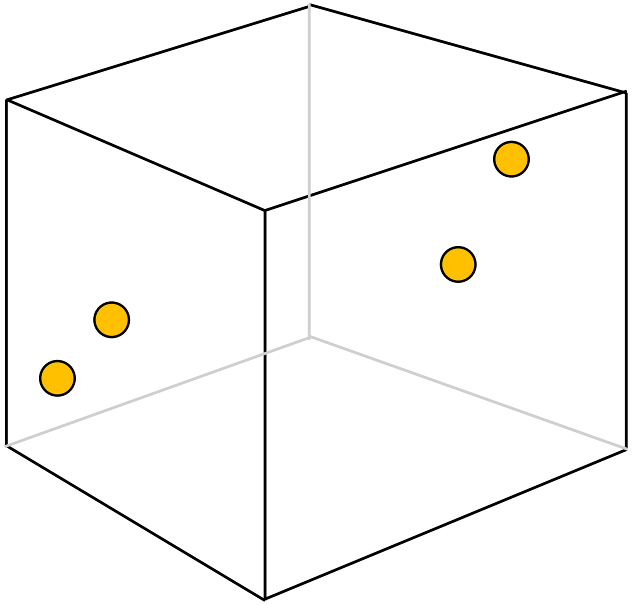
- Access range: **Entire hash table** ☹



vs.

Restricted Hashing

- Access range: **Small sub-table** ☺



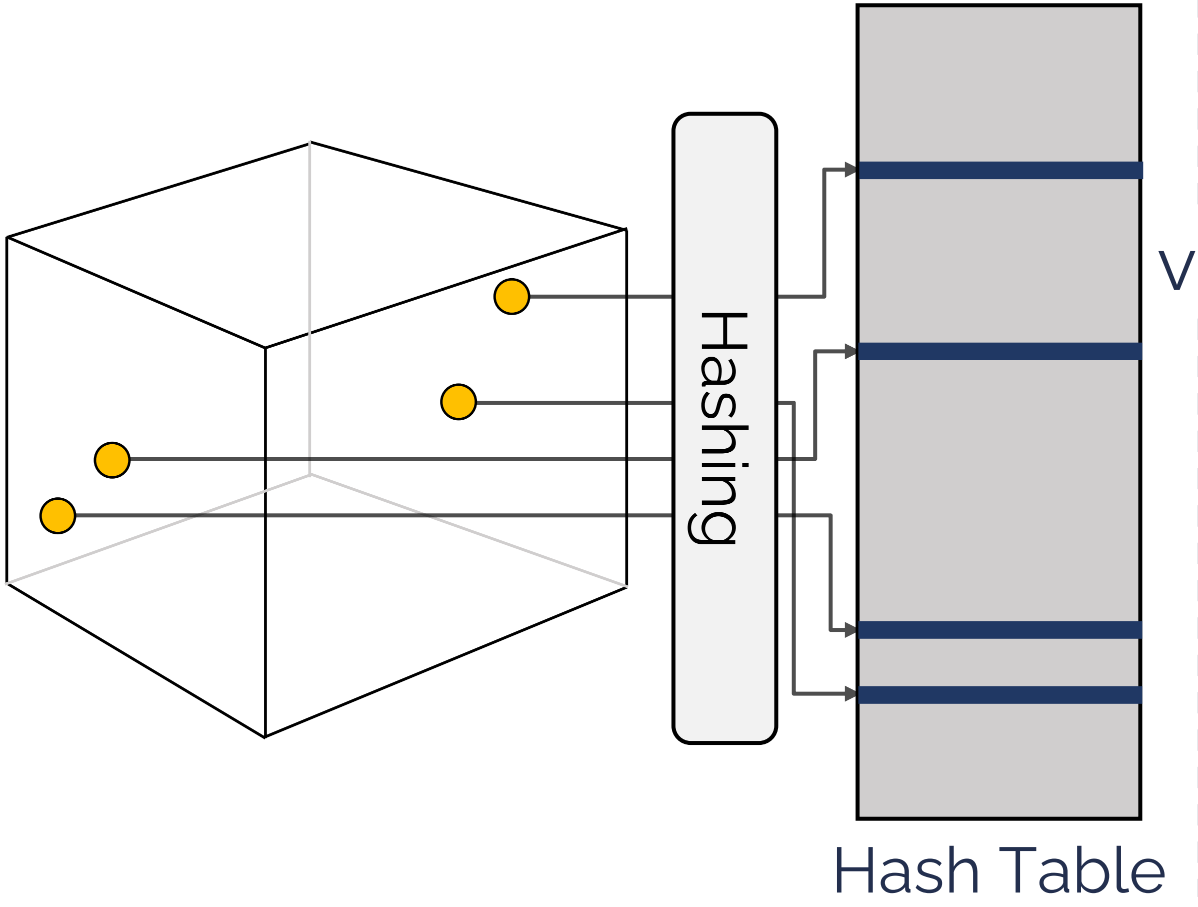
NeuRex: Hardware-Friendly Hash Encoding

Original Algorithm

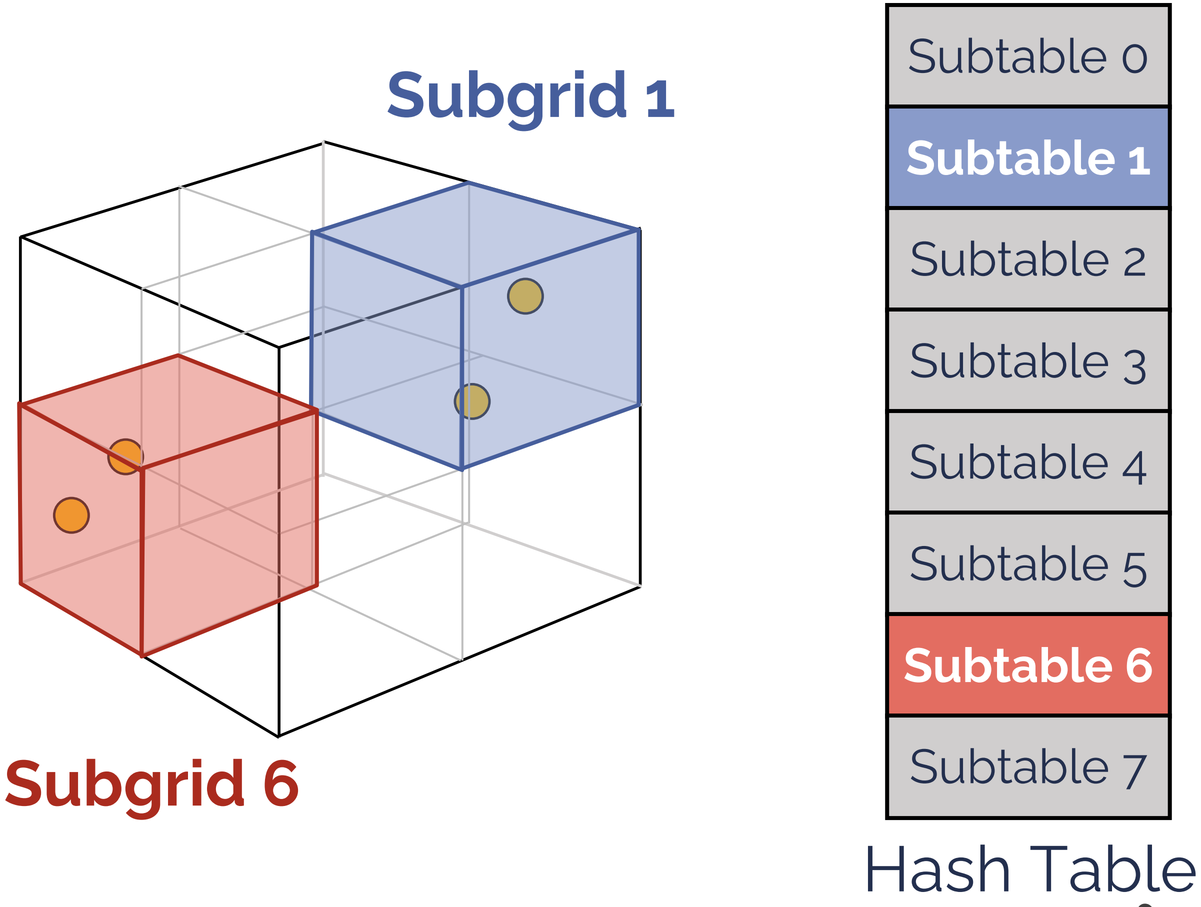
- Access range: **Entire hash table** ☹

Restricted Hashing

- Access range: **Small sub-table** ☺



VS.



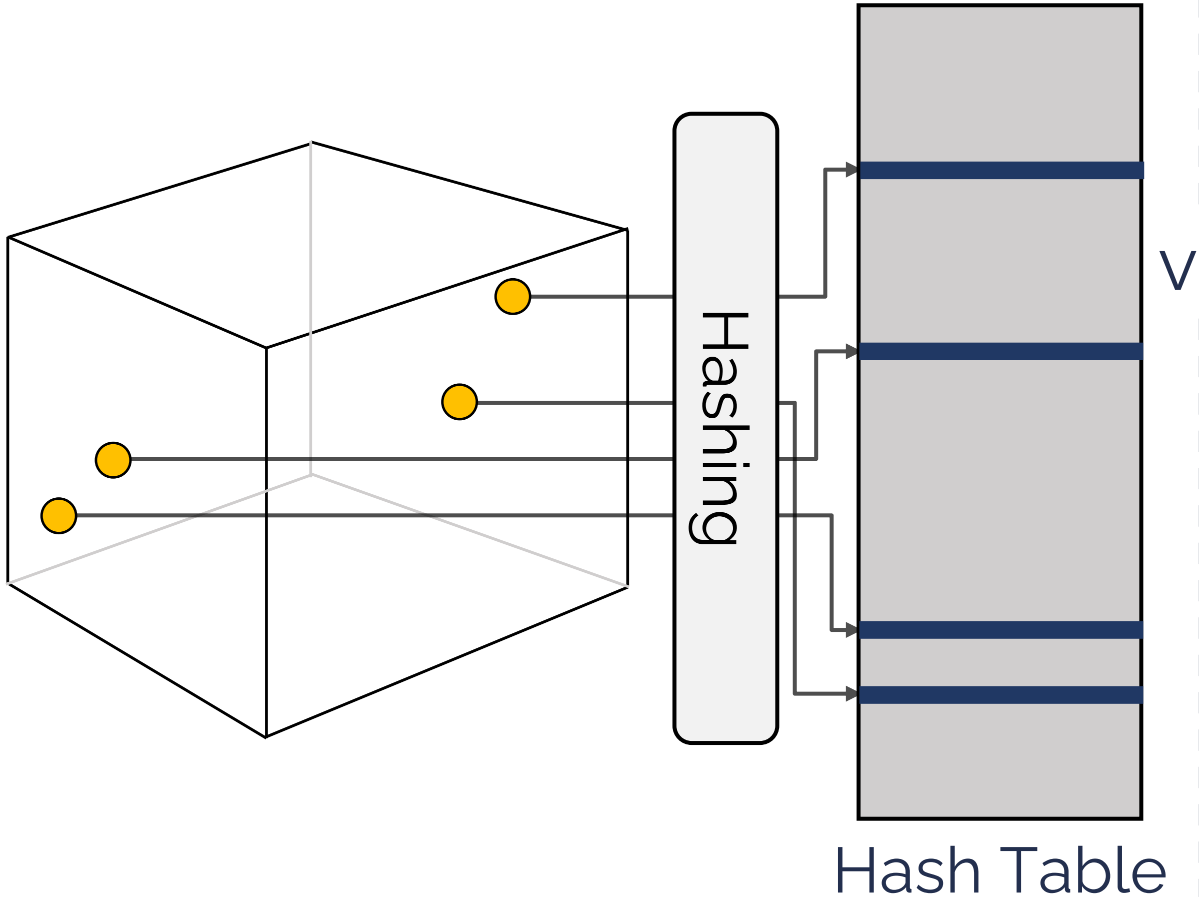
NeuRex: Hardware-Friendly Hash Encoding

Original Algorithm

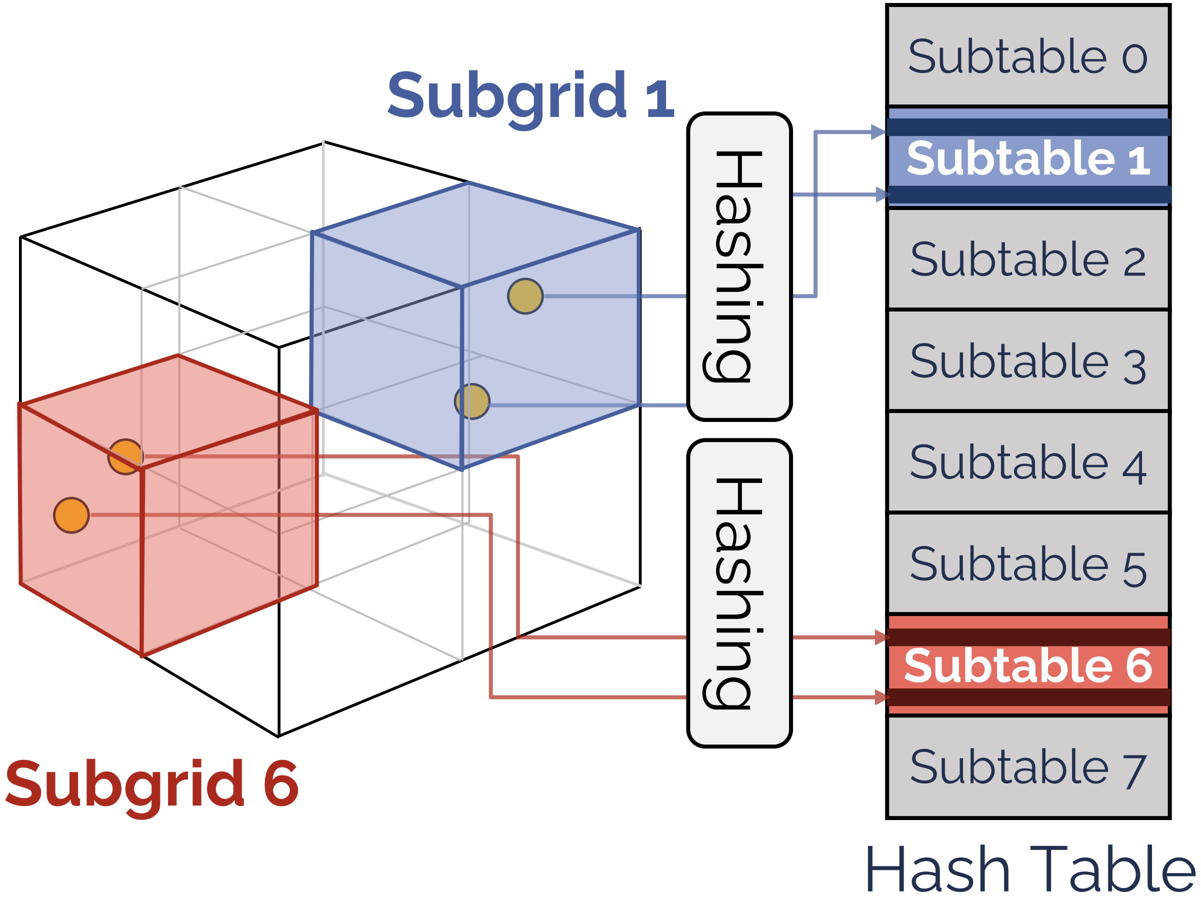
- Access range: **Entire hash table** ☹

Restricted Hashing

- Access range: **Small sub-table** ☺



vs.



NeuRex: Hardware-Friendly Hash Encoding

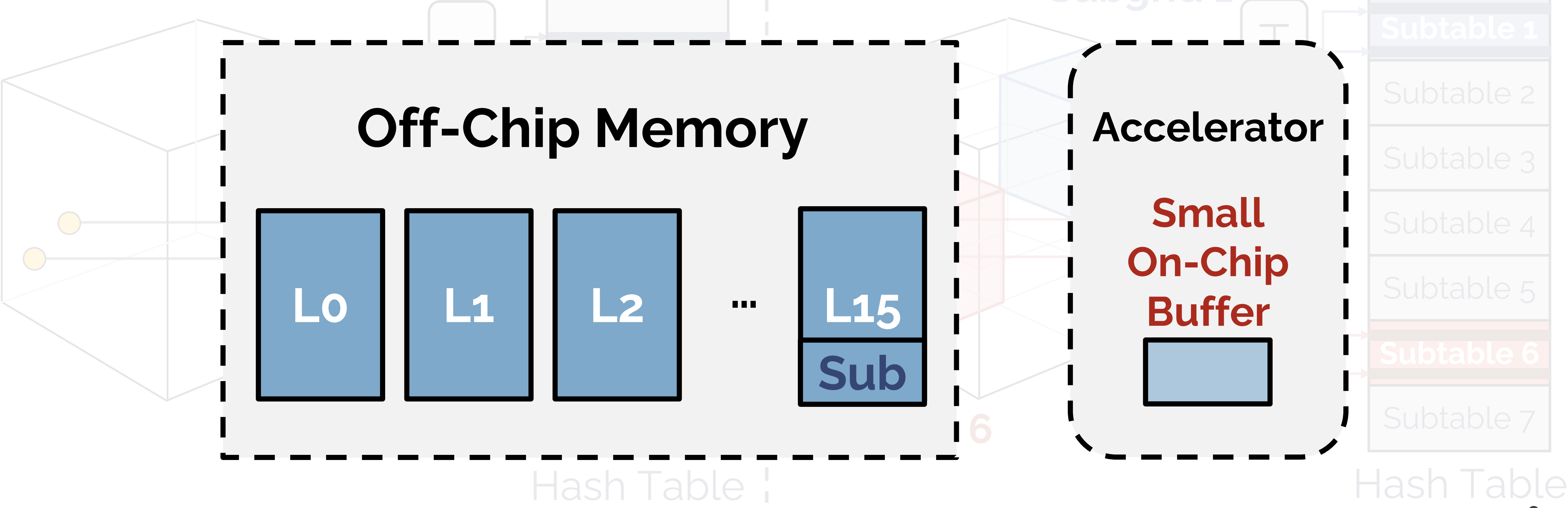
Original A

- Access

table ☺

Advantage 1. Performance Portability

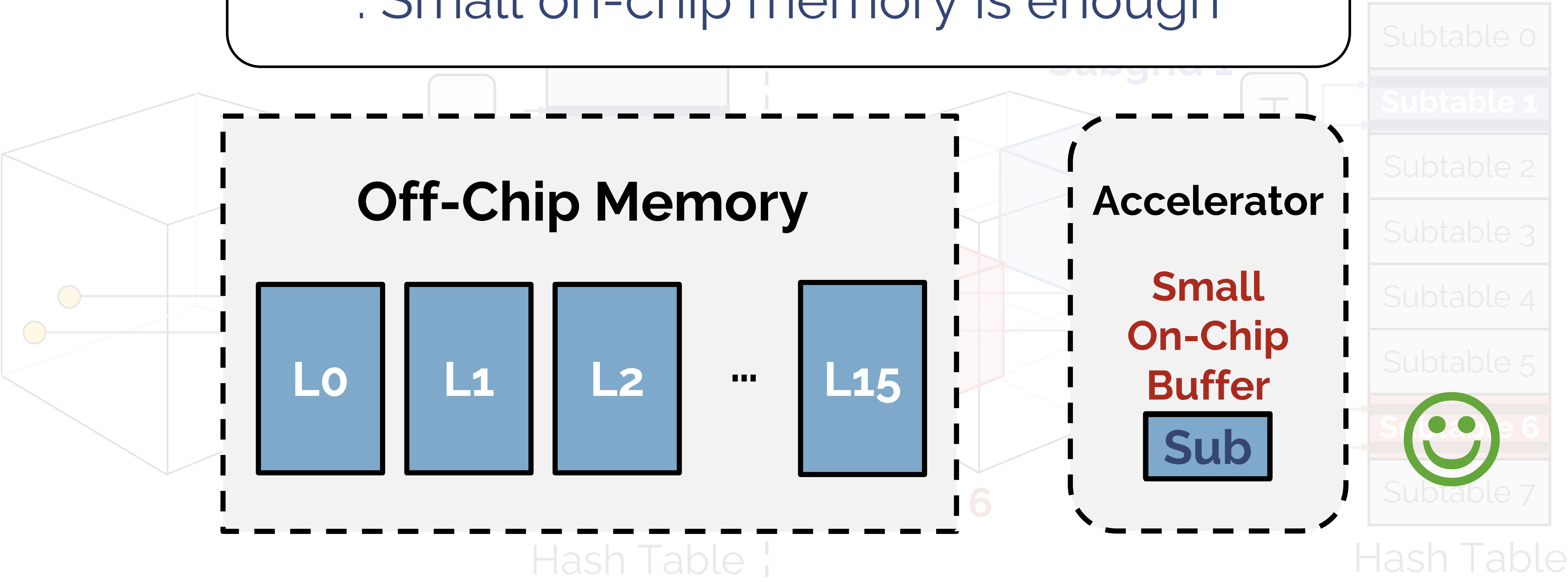
: Small on-chip memory is enough



NeuRex: Hardware-Friendly Hash Encoding

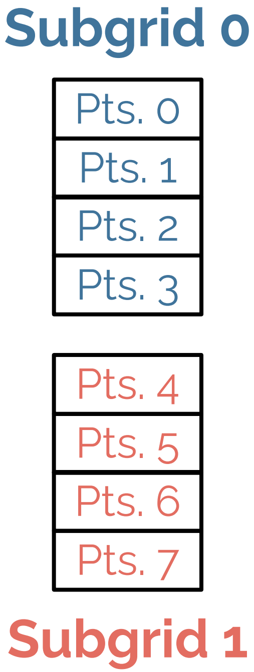
Advantage 1. Performance Portability

: Small on-chip memory is enough



Advantage 2. Parallelized Execution of ENC and MLP

Subgrid-wise execution



Subgrid 0

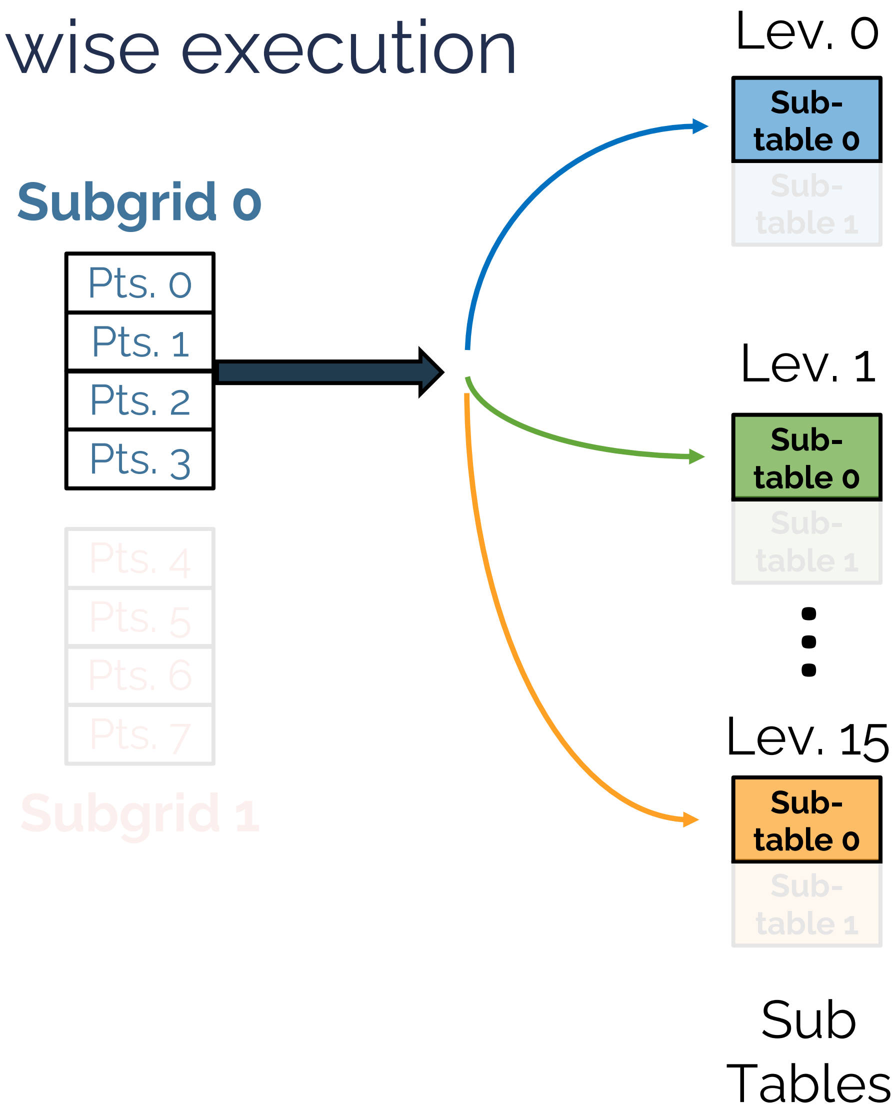
Pts. 0
Pts. 1
Pts. 2
Pts. 3

Pts. 4
Pts. 5
Pts. 6
Pts. 7

Subgrid 1

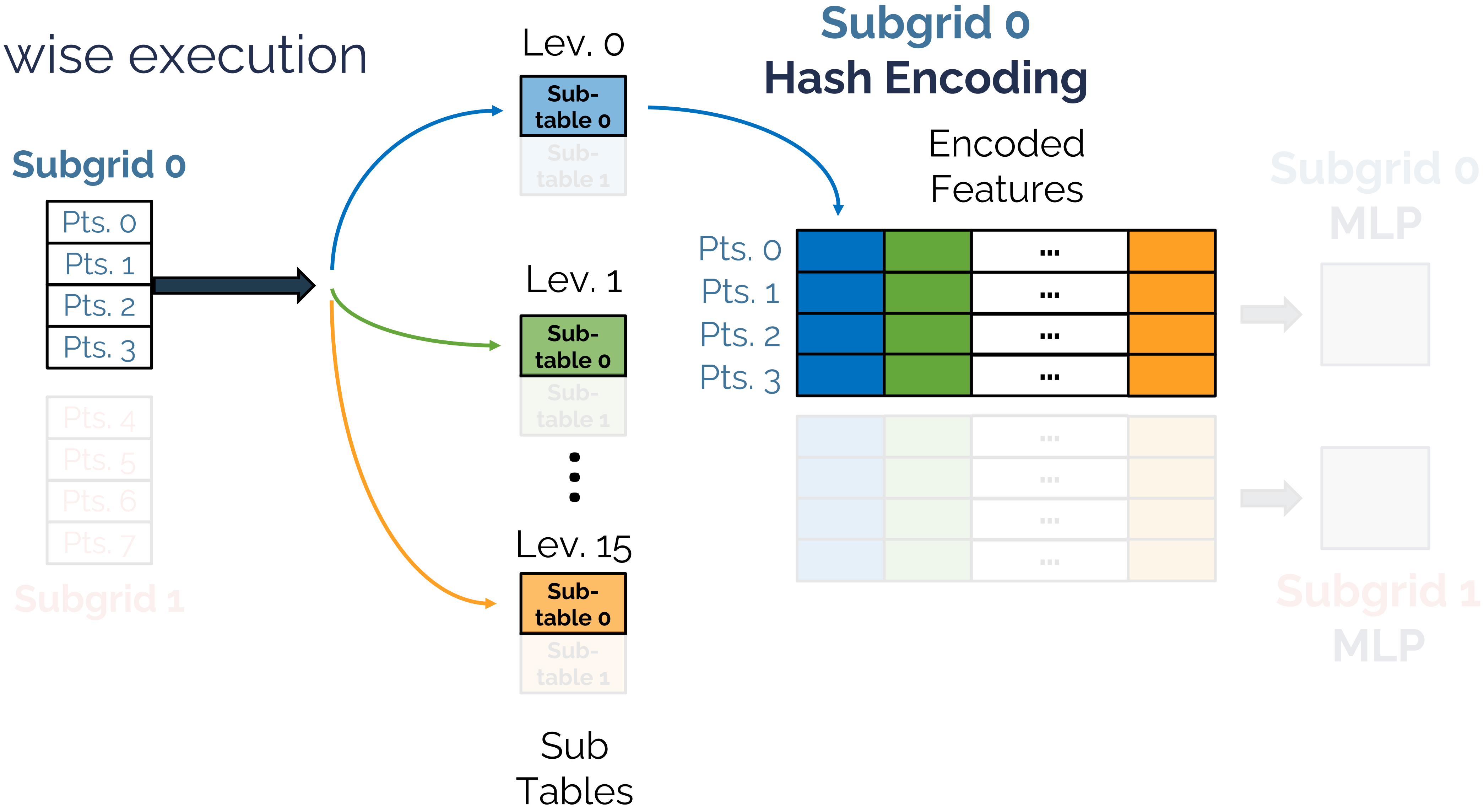
Advantage 2. Parallelized Execution of ENC and MLP

Subgrid-wise execution



Advantage 2. Parallelized Execution of ENC and MLP

Subgrid-wise execution



Subgrid 0

Pts. 0

Pts. 1

Pts. 2

Pts. 3

Pts. 4

Pts. 5

Pts. 6

Pts. 7

Subgrid 1

Lev. 0

Sub-table 0

Sub-table 1

Lev. 1

Sub-table 0

Sub-table 1

⋮

Lev. 15

Sub-table 0

Sub-table 1

Sub Tables

Subgrid 0 Hash Encoding

Encoded Features

Pts. 0 ...

Pts. 1 ...

Pts. 2 ...

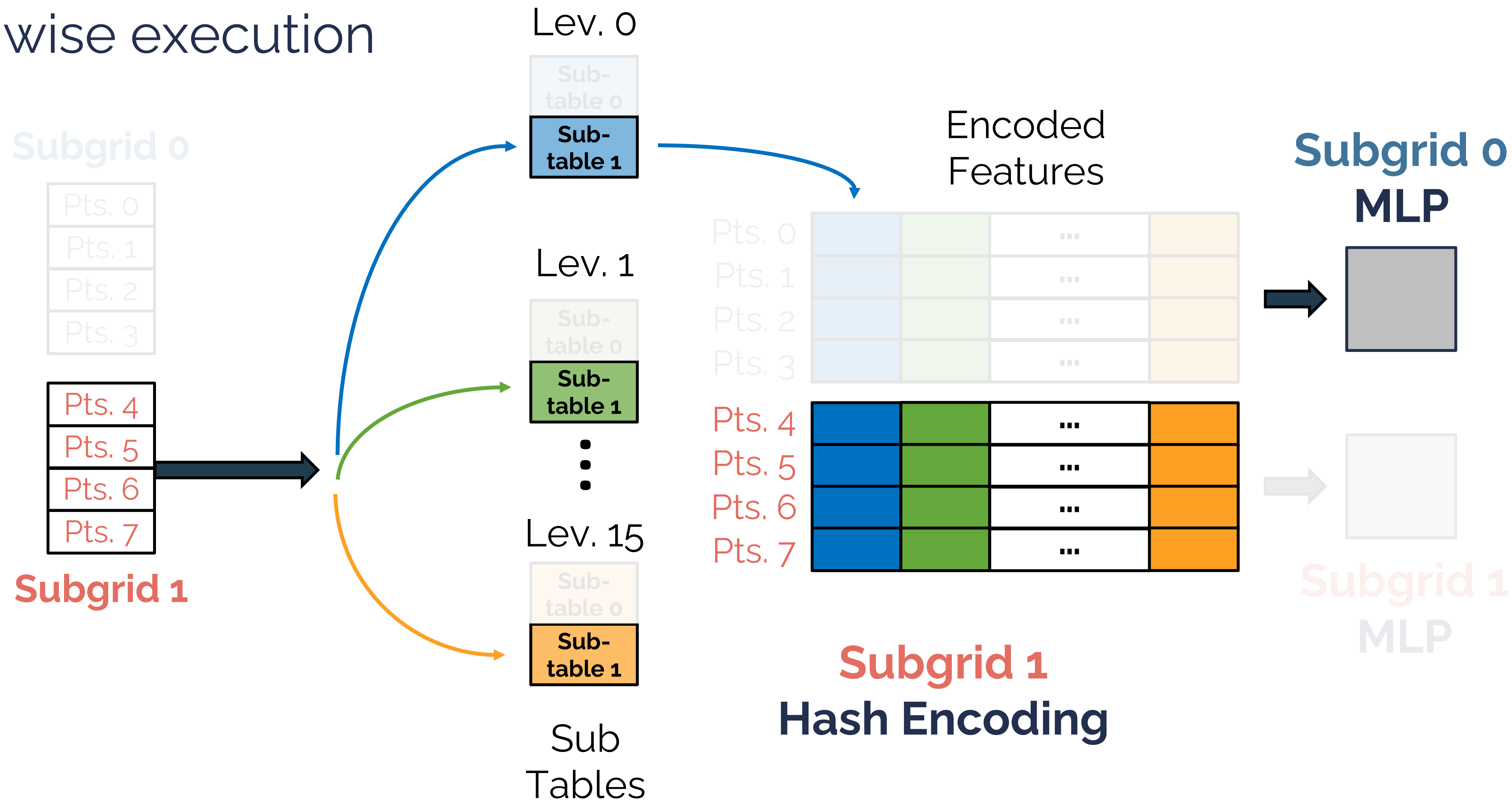
Pts. 3 ...

Subgrid 0 MLP

Subgrid 1 MLP

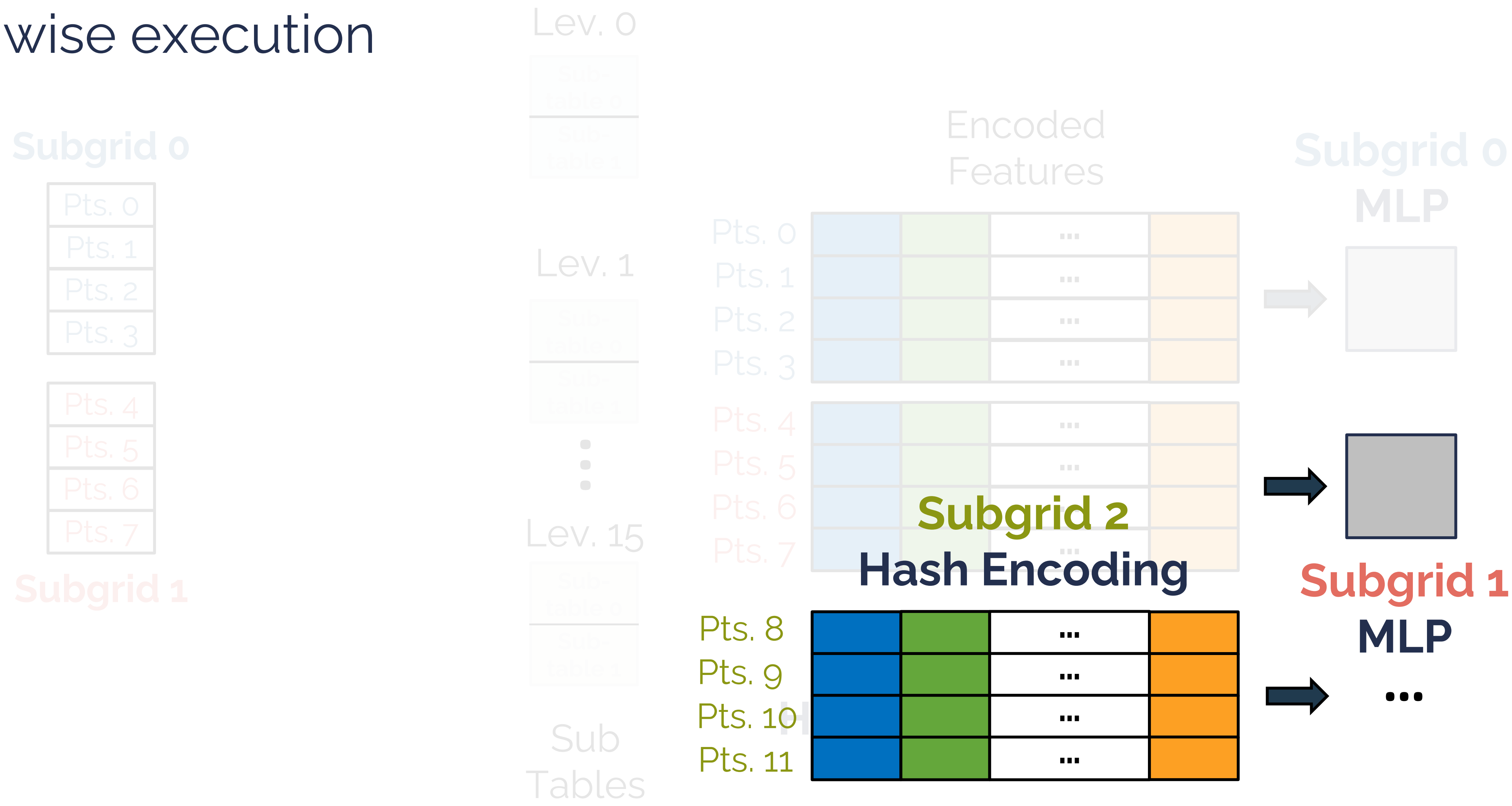
Advantage 2. Parallelized Execution of ENC and MLP

Subgrid-wise execution



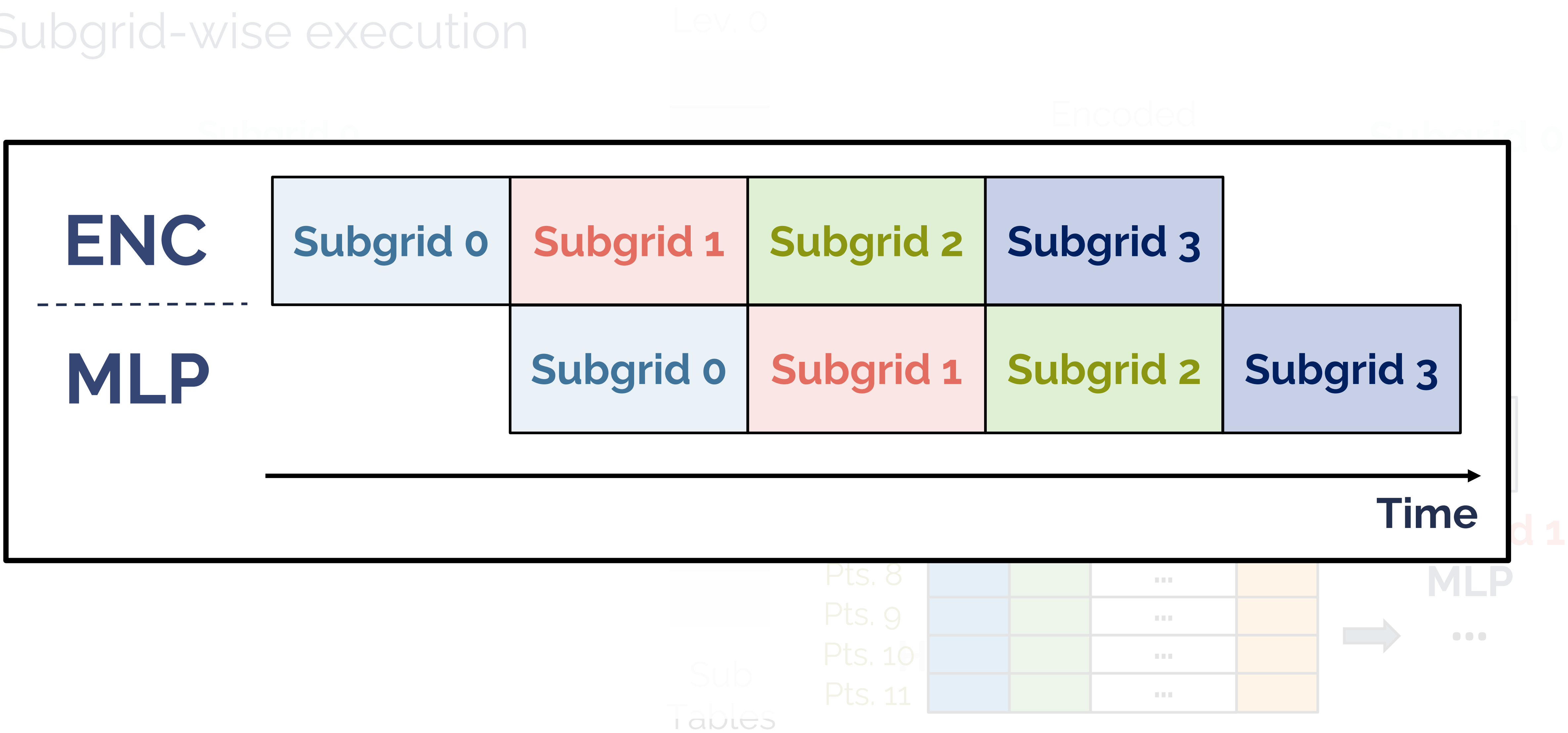
Advantage 2. Parallelized Execution of ENC and MLP

Subgrid-wise execution

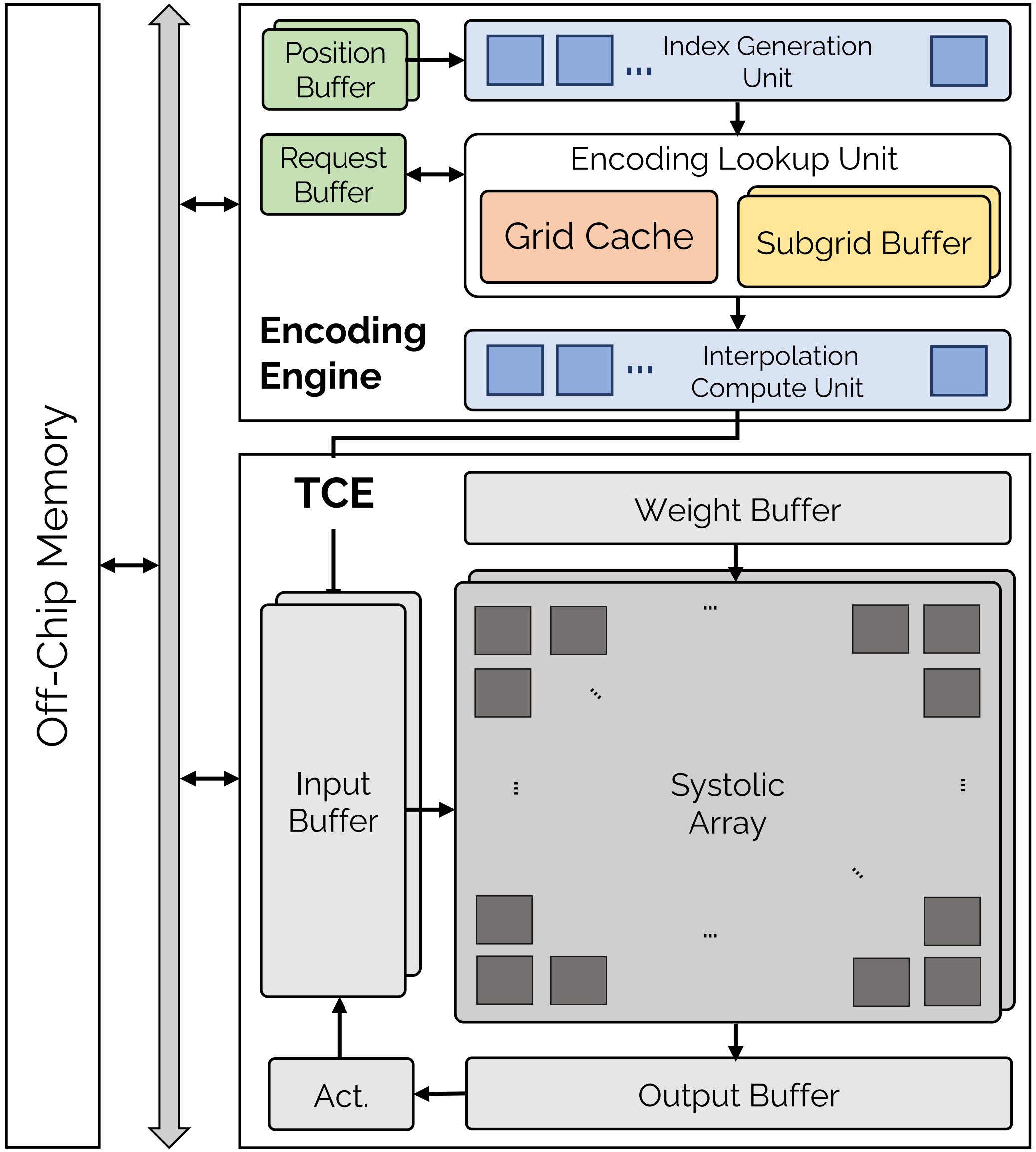


Advantage 2. Parallelized Execution of ENC and MLP

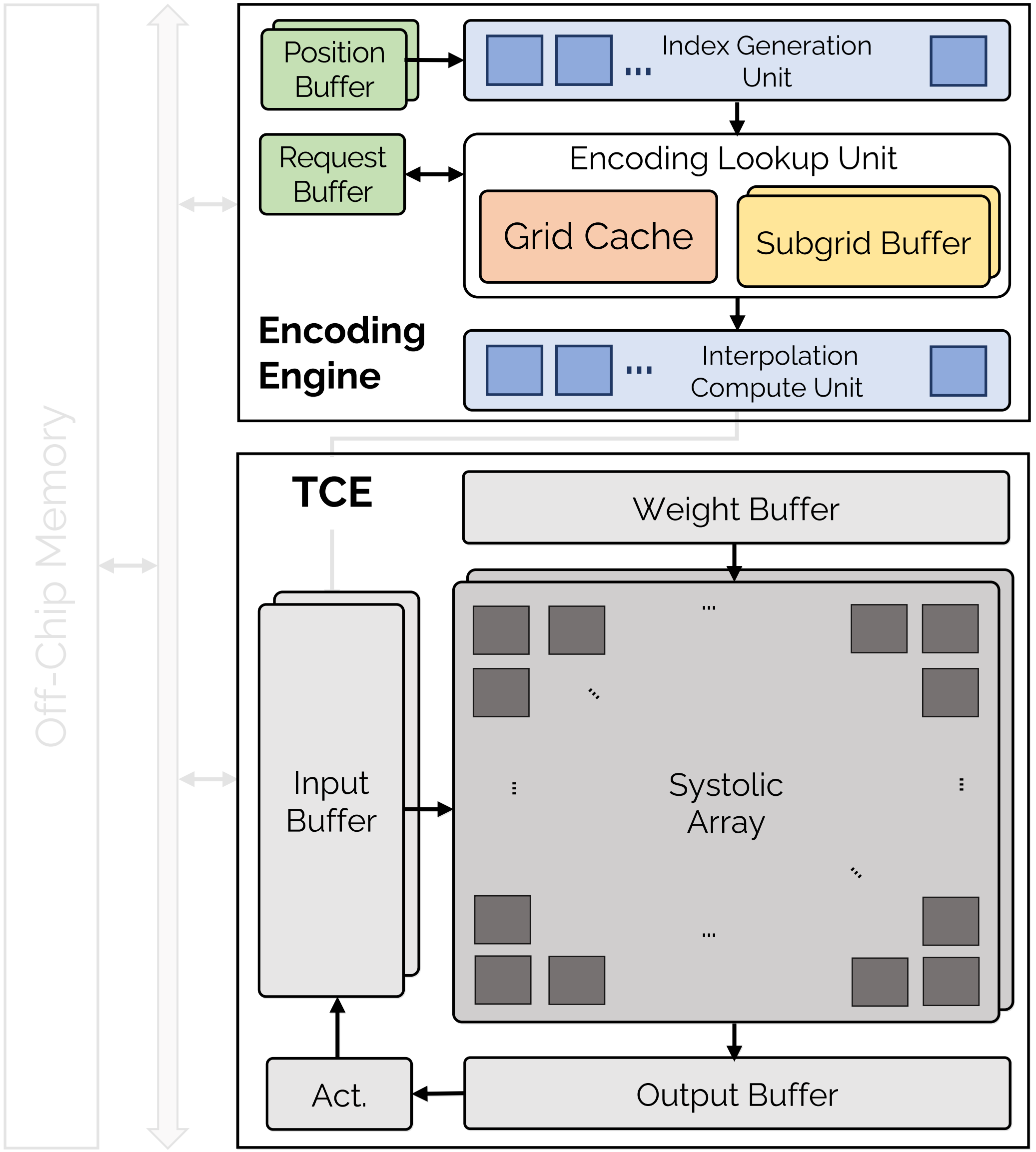
Subgrid-wise execution



NeuRex: Neural Graphics Engine



NeuRex: Neural Graphics Engine



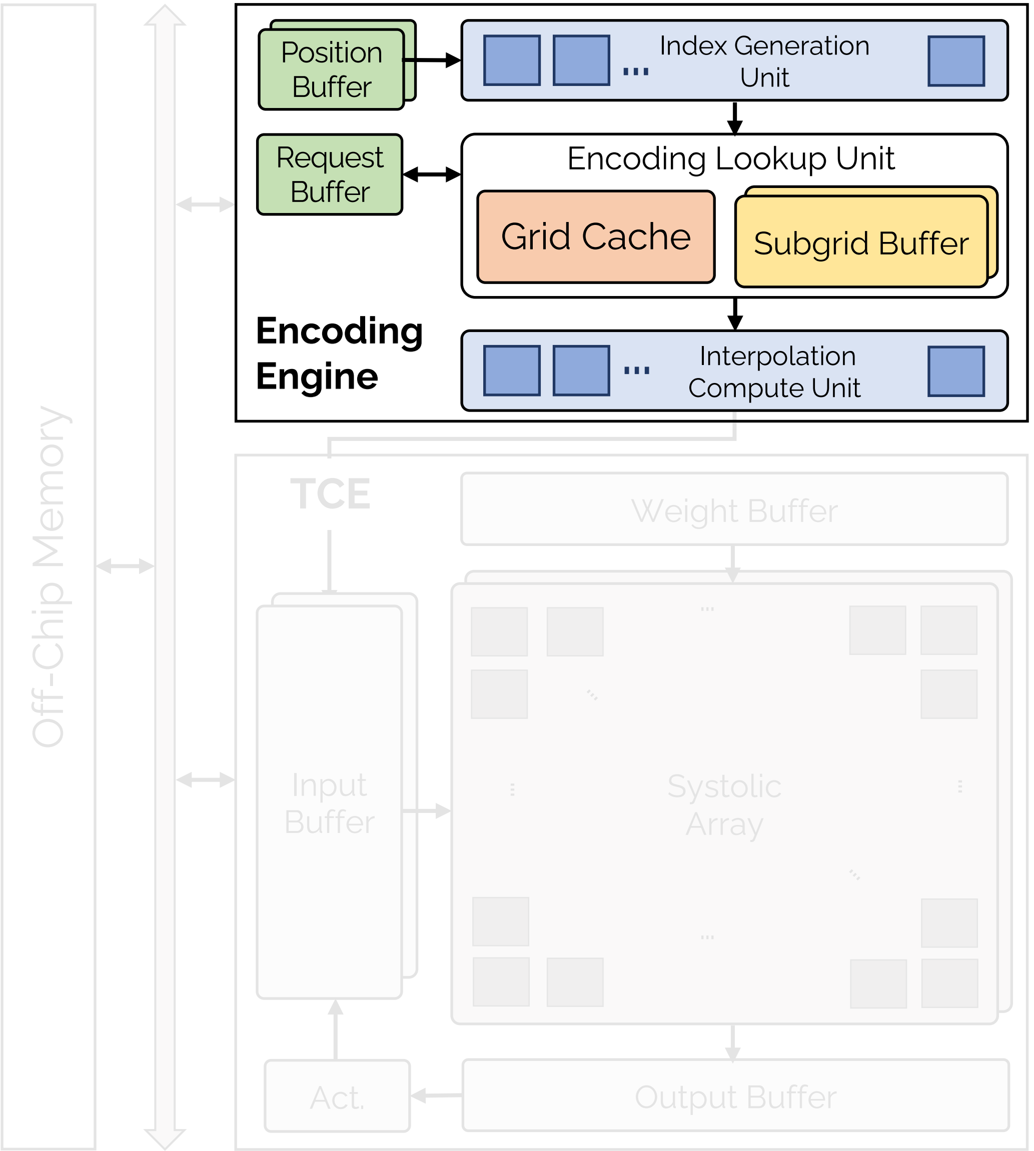
Encoding Engine (EE)

- Perform multi-resolution hash encoding

Tensor Compute Engine (TCE)

- Perform MLP execution

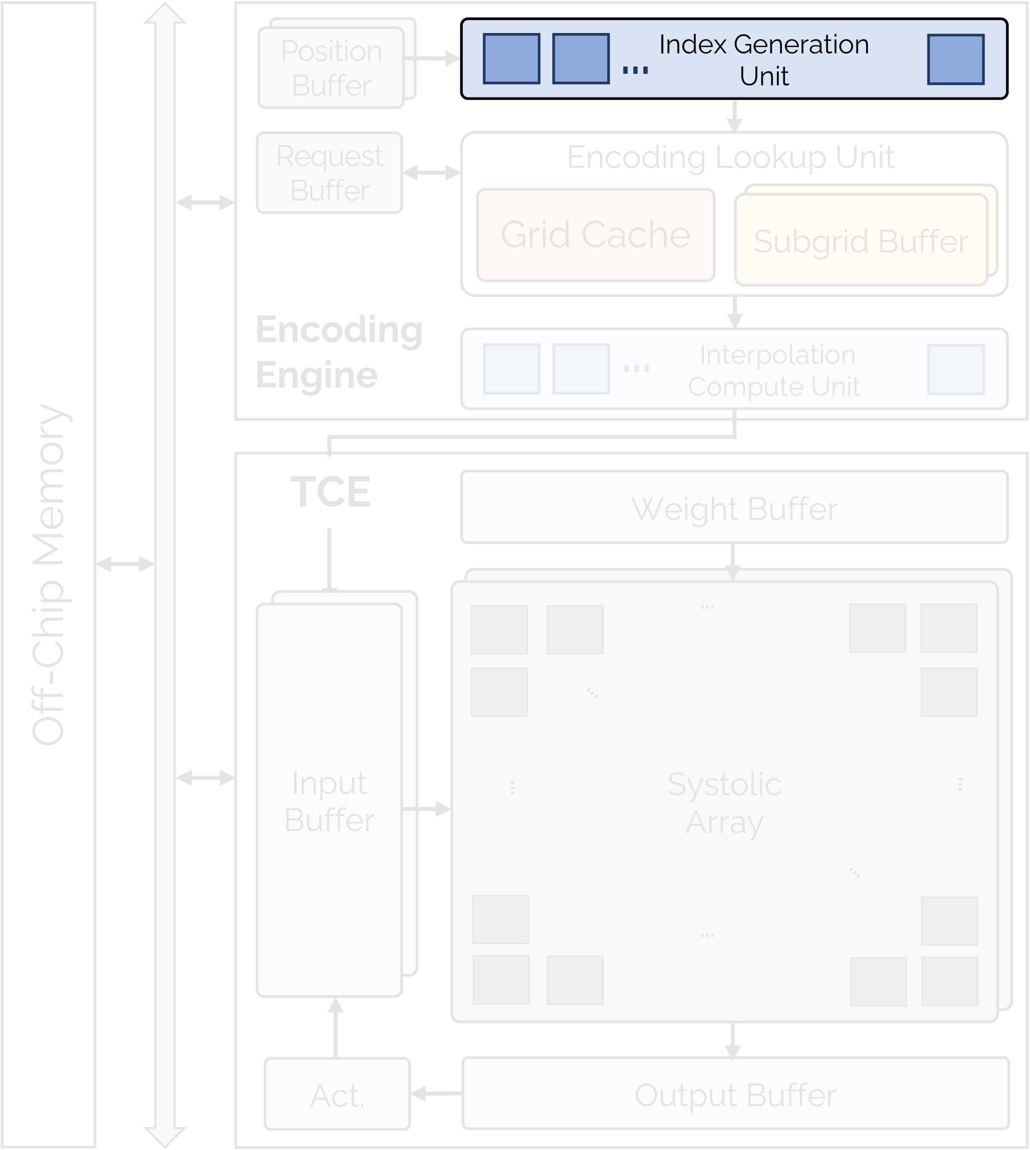
NeuRex: Neural Graphics Engine



Encoding Engine (EE)

- Perform multi-resolution hash encoding

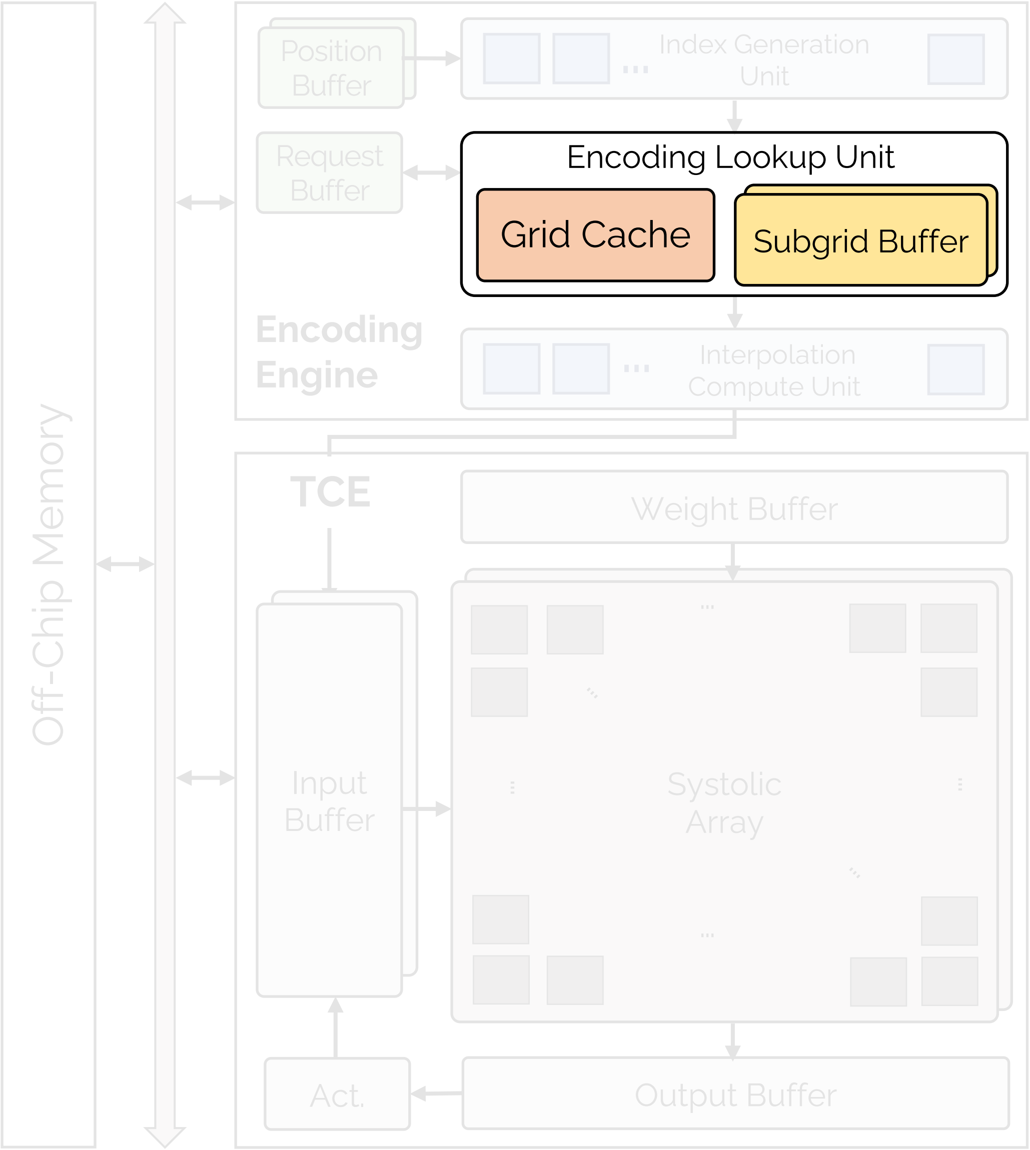
NeuRex: Neural Graphics Engine



Index Generation Unit (IGU)

- Calculate hash indices
- Compute interpolation weights

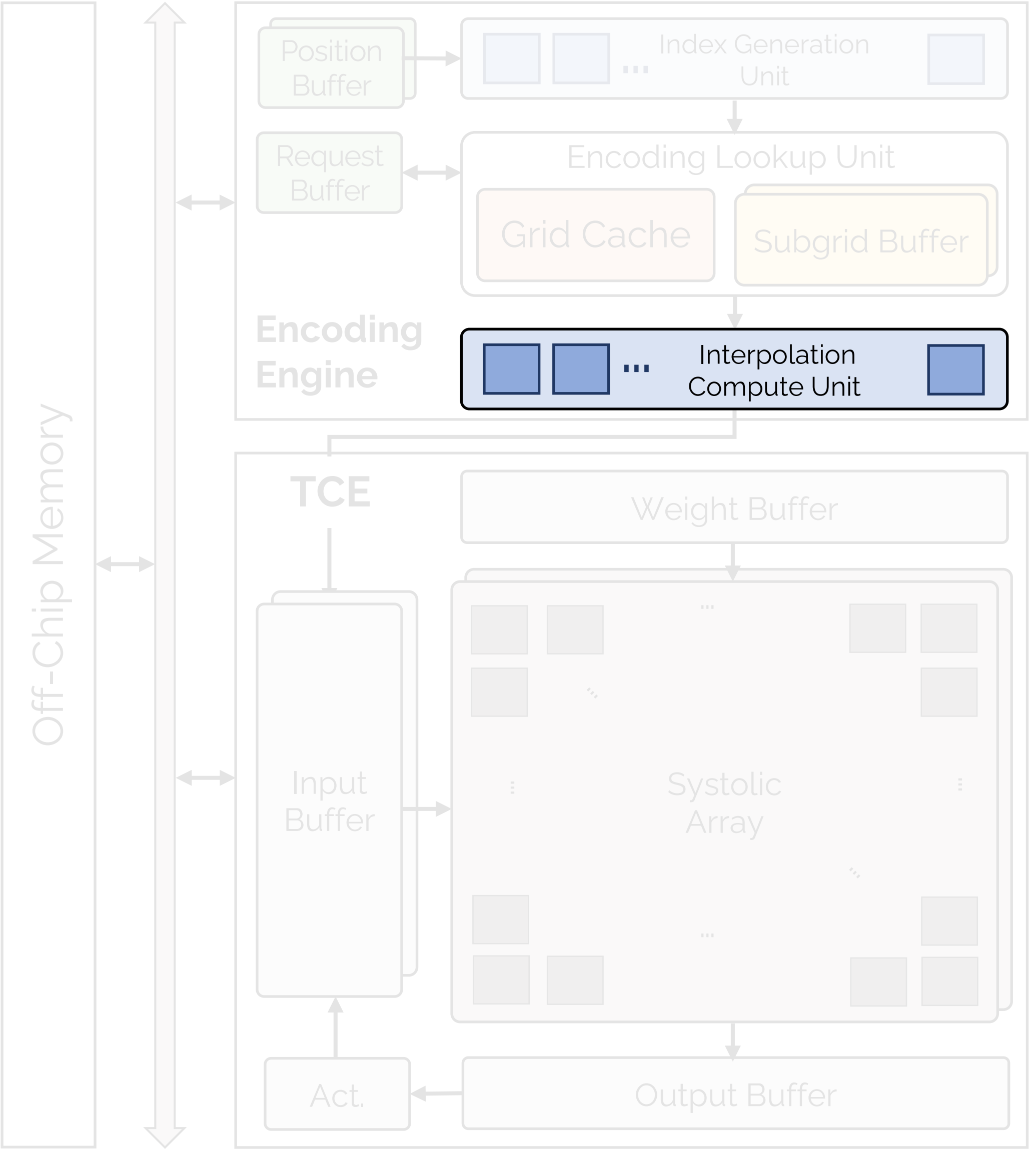
NeuRex: Neural Graphics Engine



Encoding Lookup Unit (ELU)

- Fetch hash table entries from on-chip buffers

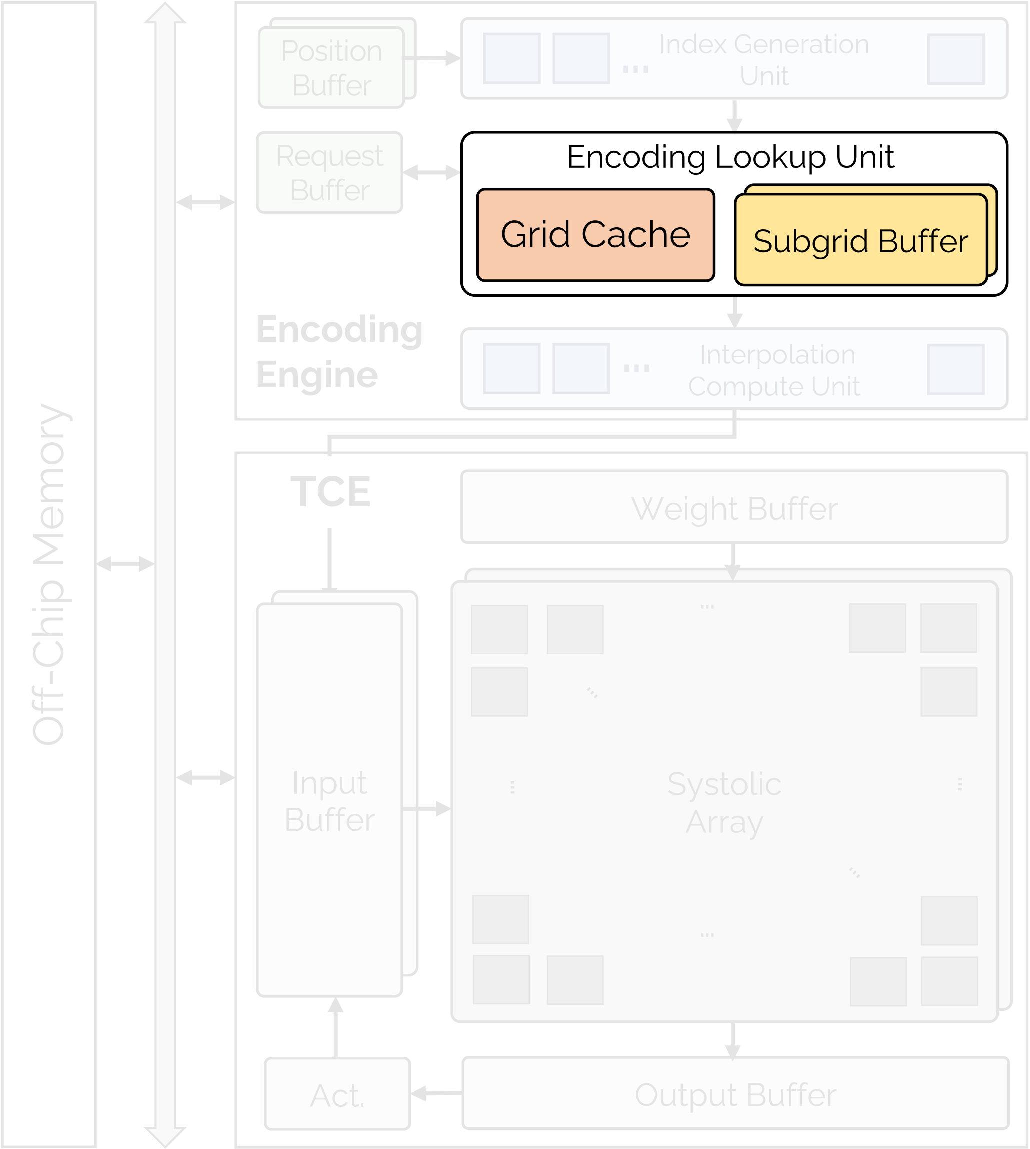
NeuRex: Neural Graphics Engine



Interpolation Compute Unit (ICU)

- Interpolate 8 hash entries

NeuRex: Neural Graphics Engine



Encoding Lookup Unit (ELU)

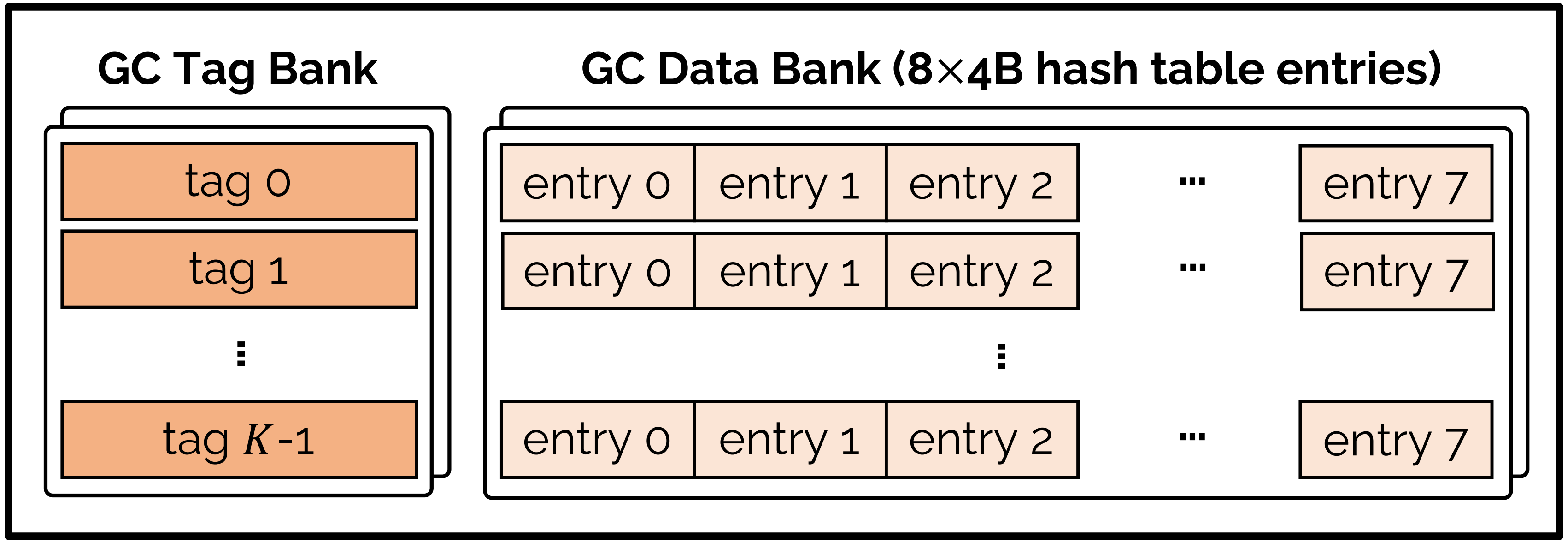
- Fetch hash table entries from on-chip buffers

Encoding Lookup Unit

Encoding Lookup Unit

Grid Cache

- A cache for **coarse levels**’ hash encoding (L=0,1,2,…)
- Cache block: Coalesced 8 hash entries



GC Tag Bank

tag 0
tag 1
⋮
tag K-1

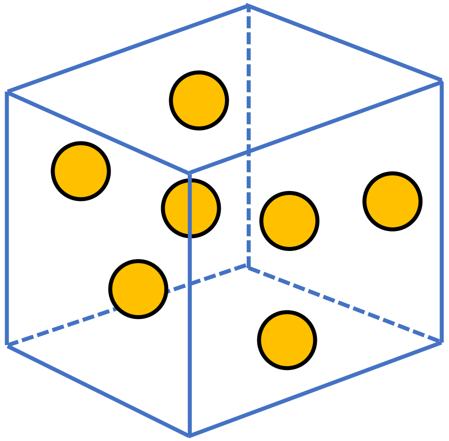
GC Data Bank (8×4B hash table entries)

entry 0	entry 1	entry 2	···	entry 7
entry 0	entry 1	entry 2	···	entry 7
⋮				
entry 0	entry 1	entry 2	···	entry 7

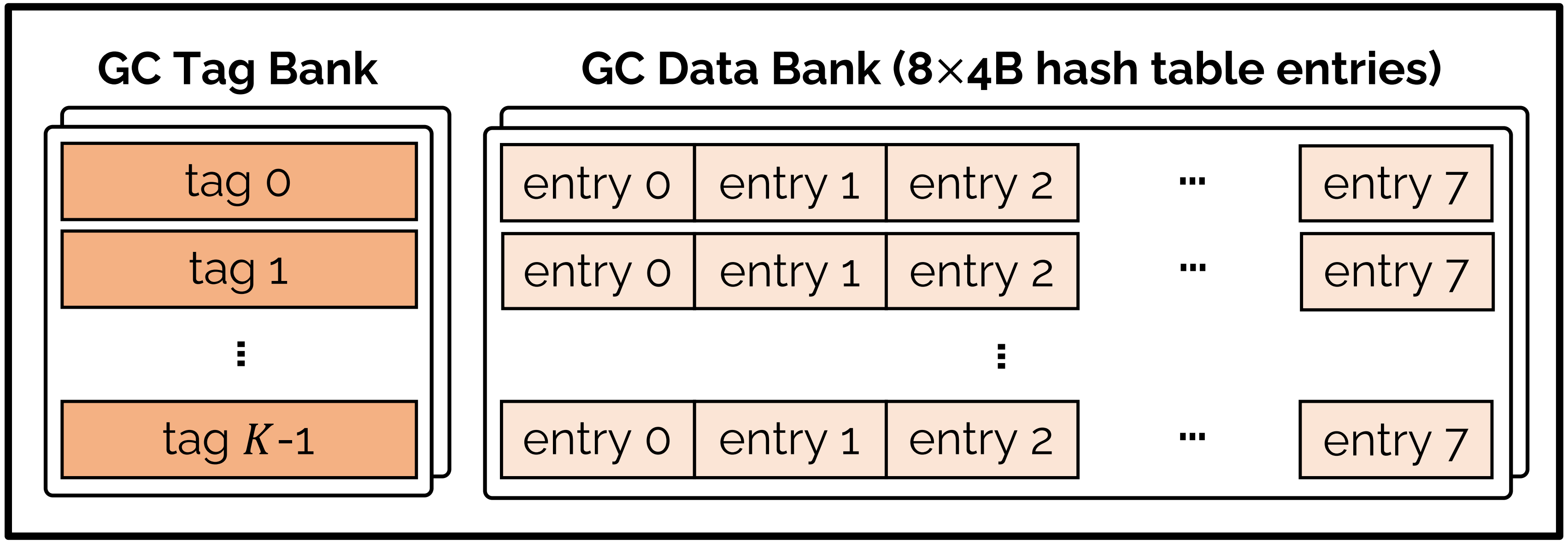
Encoding Lookup Unit

Grid Cache

- A cache for **coarse levels**' hash encoding (L=0,1,2,…)
- Cache block: Coalesced 8 hash entries



Coarse Level



GC Tag Bank

tag 0

tag 1

⋮

tag K-1

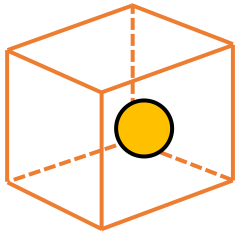
GC Data Bank (8×4B hash table entries)

entry 0 | entry 1 | entry 2 | ⋯ | entry 7

entry 0 | entry 1 | entry 2 | ⋯ | entry 7

⋮

entry 0 | entry 1 | entry 2 | ⋯ | entry 7

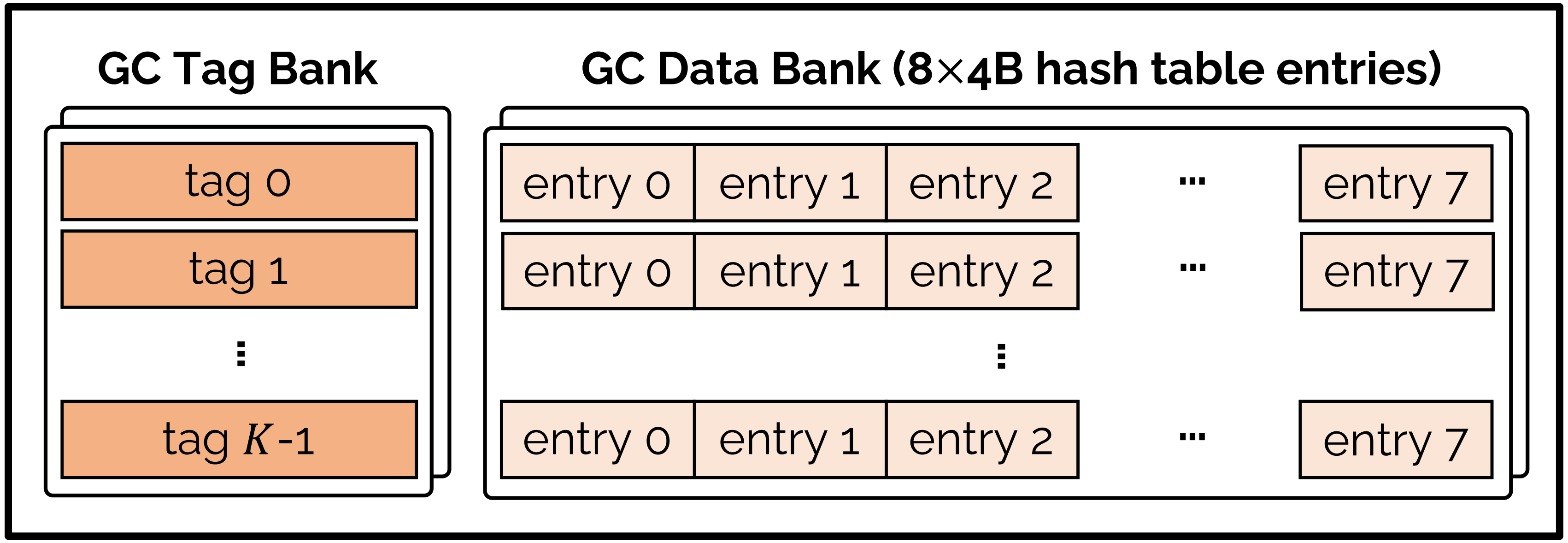


Fine Level

Encoding Lookup Unit

Grid Cache

- A cache for **coarse levels**’ hash encoding (L=0,1,2,…)
- Cache block: Coalesced 8 hash entries

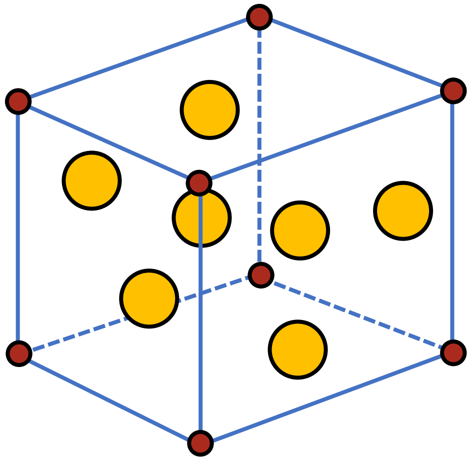


GC Tag Bank

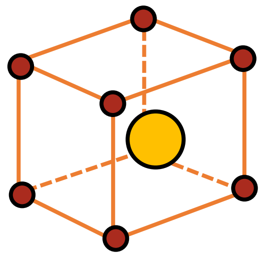
tag 0
tag 1
⋮
tag K-1

GC Data Bank (8×4B hash table entries)

entry 0 | entry 1 | entry 2 | … | entry 7
entry 0 | entry 1 | entry 2 | … | entry 7
⋮
entry 0 | entry 1 | entry 2 | … | entry 7



Coarse Level



Fine Level

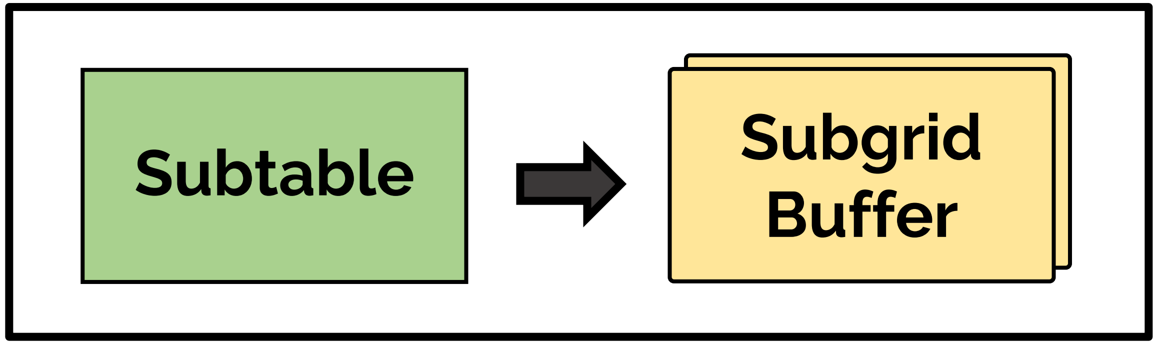
Encoding Lookup Unit

Grid Cache

- A cache for **coarse levels**' hash encoding (L=0,1,2,…)
- Cache block: Coalesced 8 hash entries

Subgrid Buffer

- A buffer for **fine levels**' hash encoding (L=15,14,13,…)
- **Restricted hashing** reduces the required on-chip memory size for hash encoding! ☺



Outline

- **Background**
  - Neural Rendering
  - Multi-resolution Hash Encoding
- **I-NGP Optimization & Limitations**
- **NeuRex: Efficient Neural Rendering Accelerator**
  - Restricted Hashing
  - Neural Graphics Engine
- **Evaluation**
- **Conclusion**

Methodology

RTL Implementation

- Process node: 28nm technology

Performance Evaluation

- Cycle-level simulator

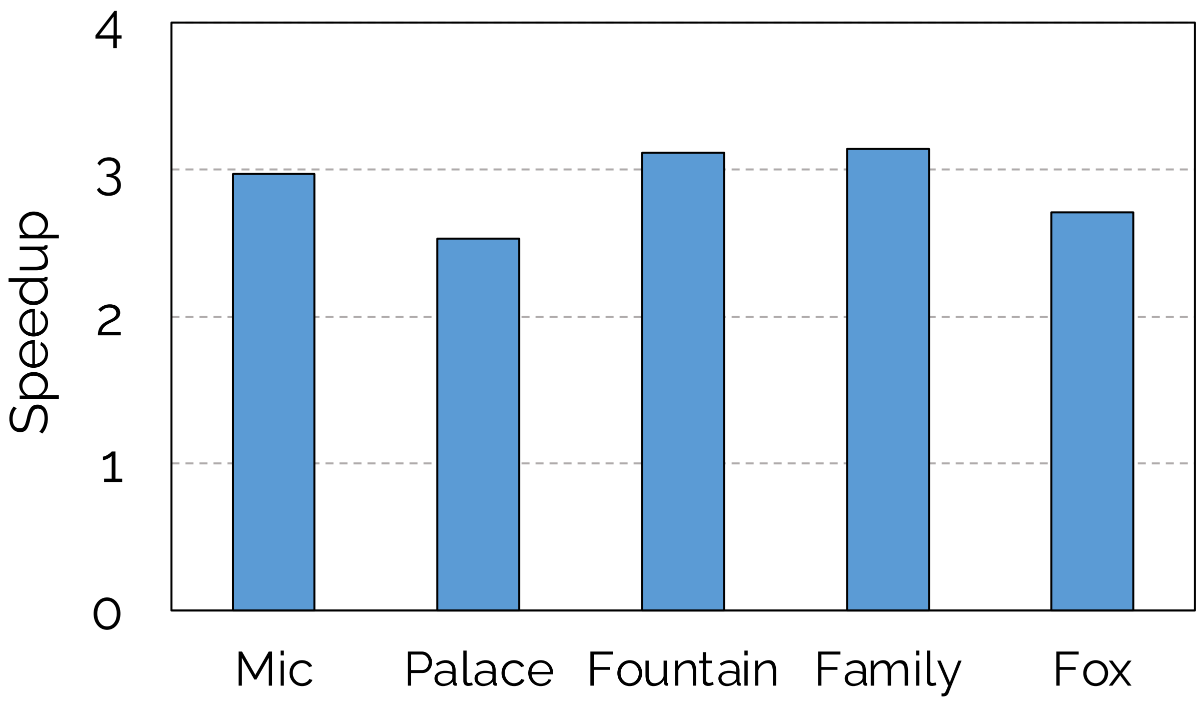
Hardware Variants & Baselines

- NeuRex-Server <-> RTX 3070
  - Area: **21.37mm²** <-> 392.5mm²
- NeuRex-Edge <-> Xavier NX
  - Area: **3.14mm²** <-> 350mm²

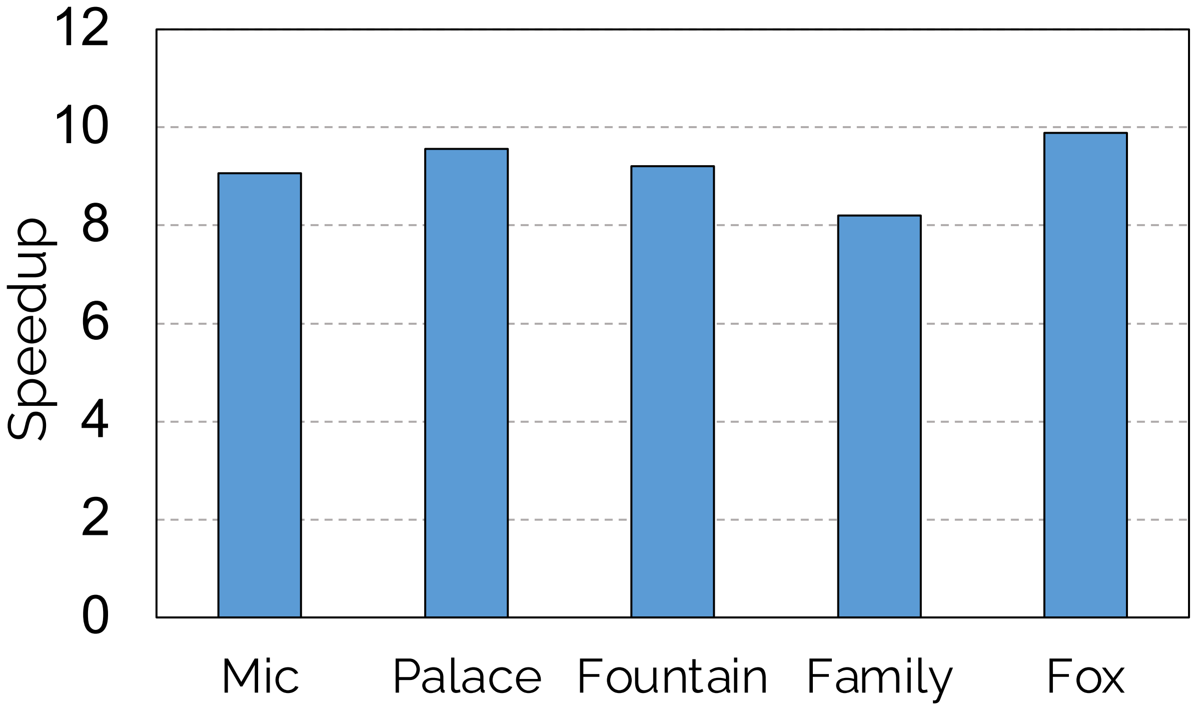
Evaluated Workloads

Dataset	Scene (Resolution)	Type
Syn-NeRF	Mic (800x800)	Synthetic
Syn-NSVF	Palace (800x800)	
BlendedMVS	Fountain (768x576)	Real world
Tanks& Temples	Family (1920x1080)	
Instant-NGP	Fox (1080x1920)	

Performance: NeRF

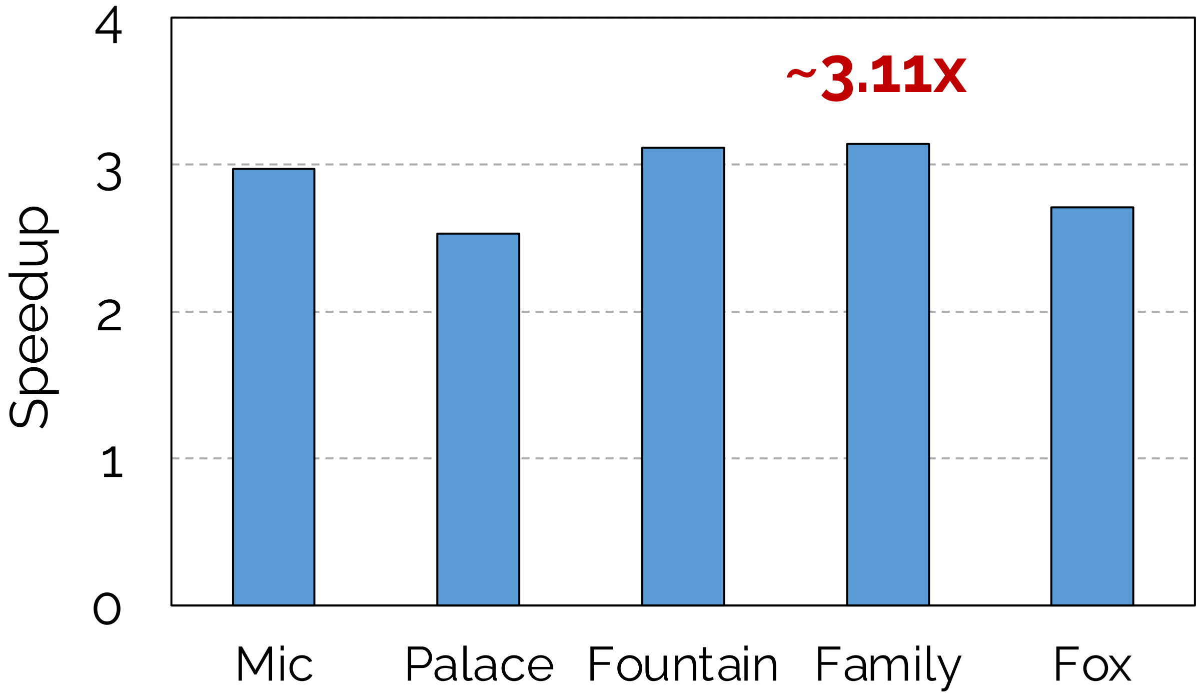


NeuRex-Server

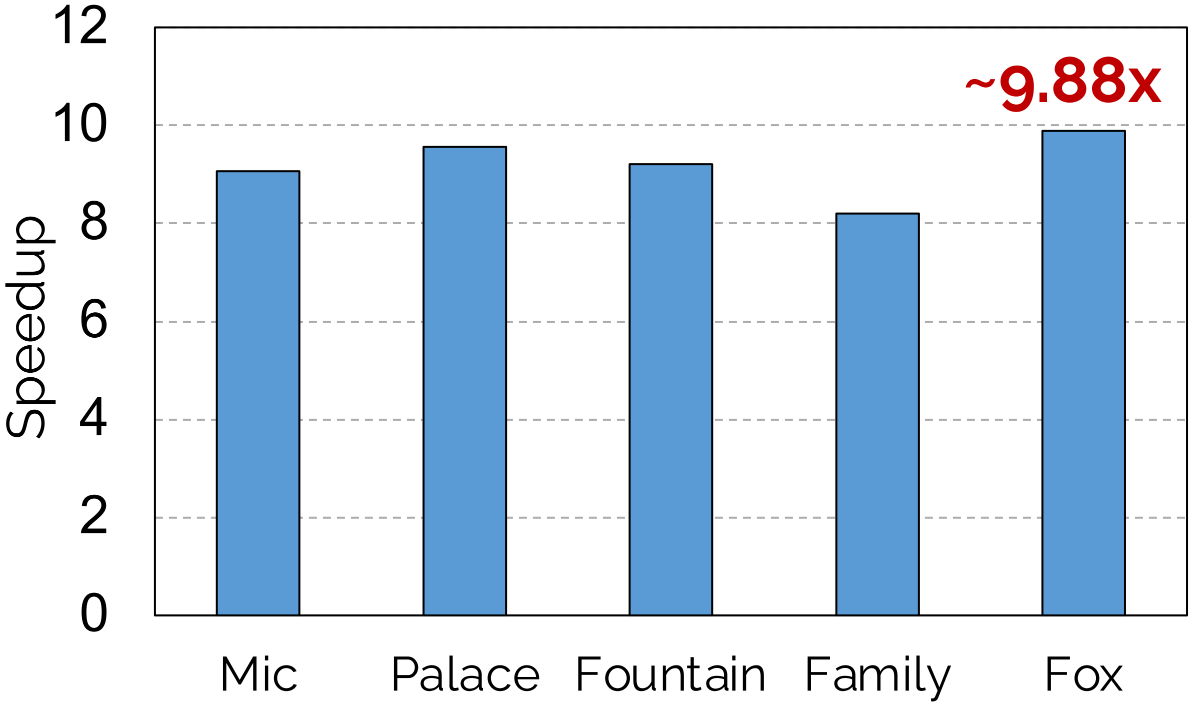


NeuRex-Edge

Performance: NeRF

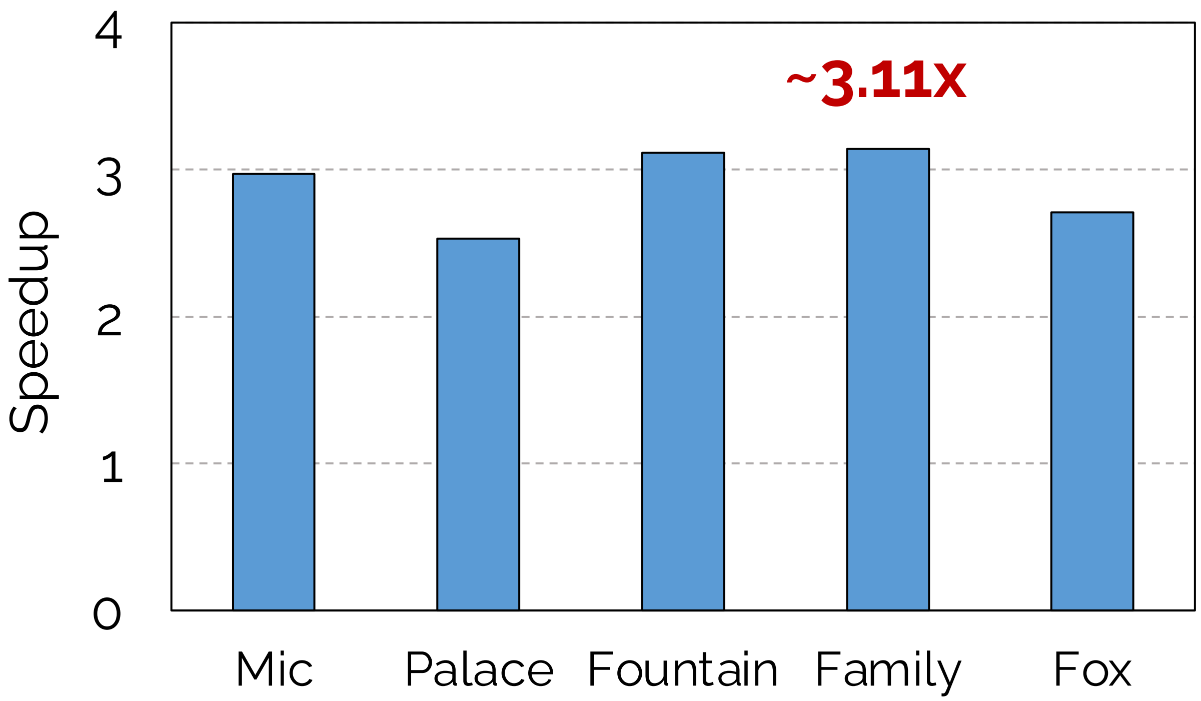


NeuRex-Server

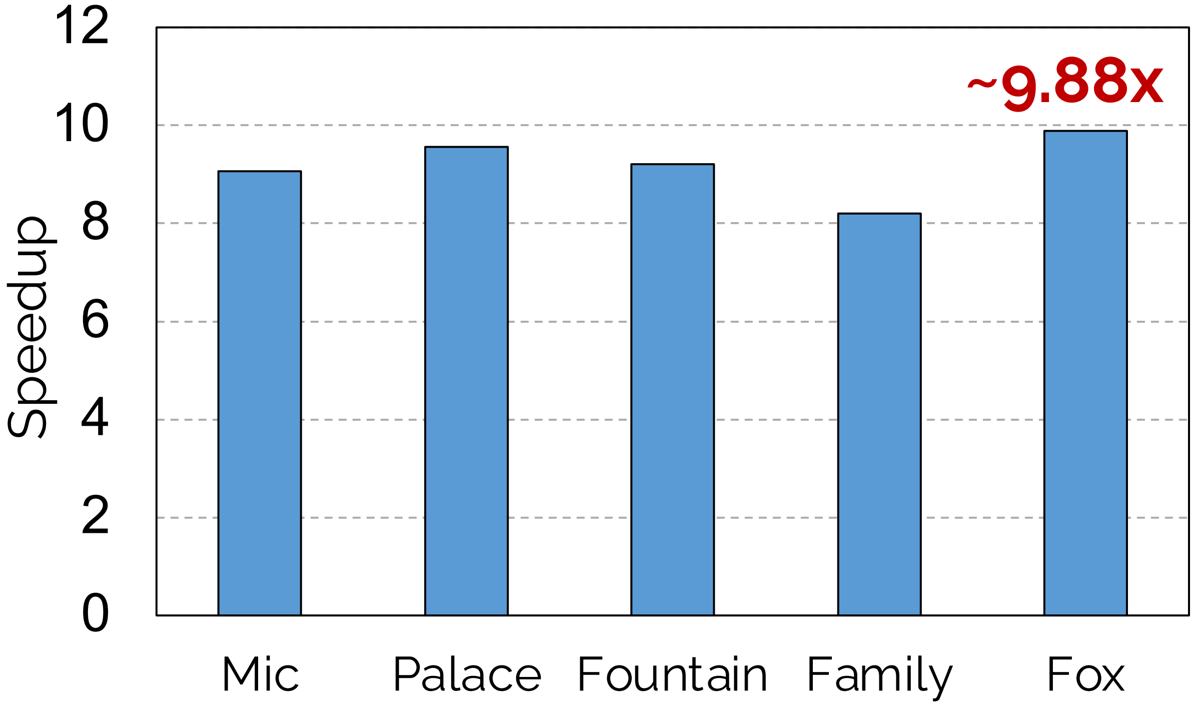


NeuRex-Edge

Performance: NeRF



NeuRex-Server

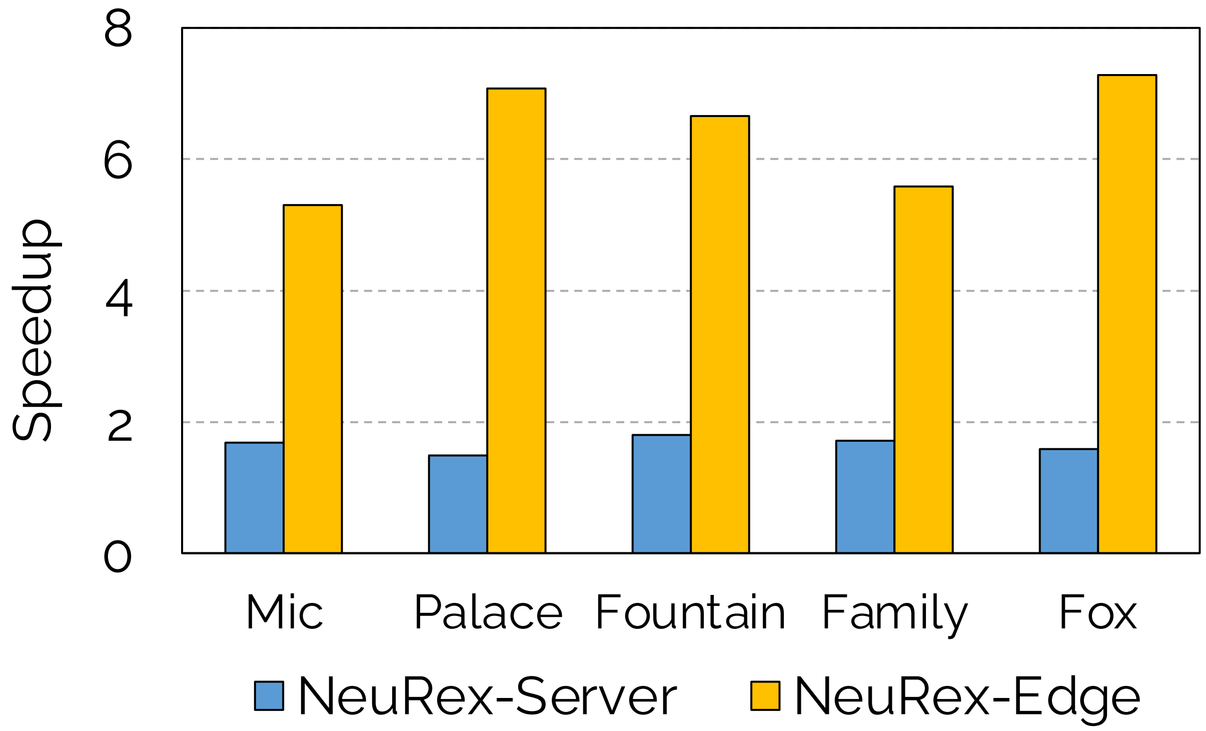


NeuRex-Edge

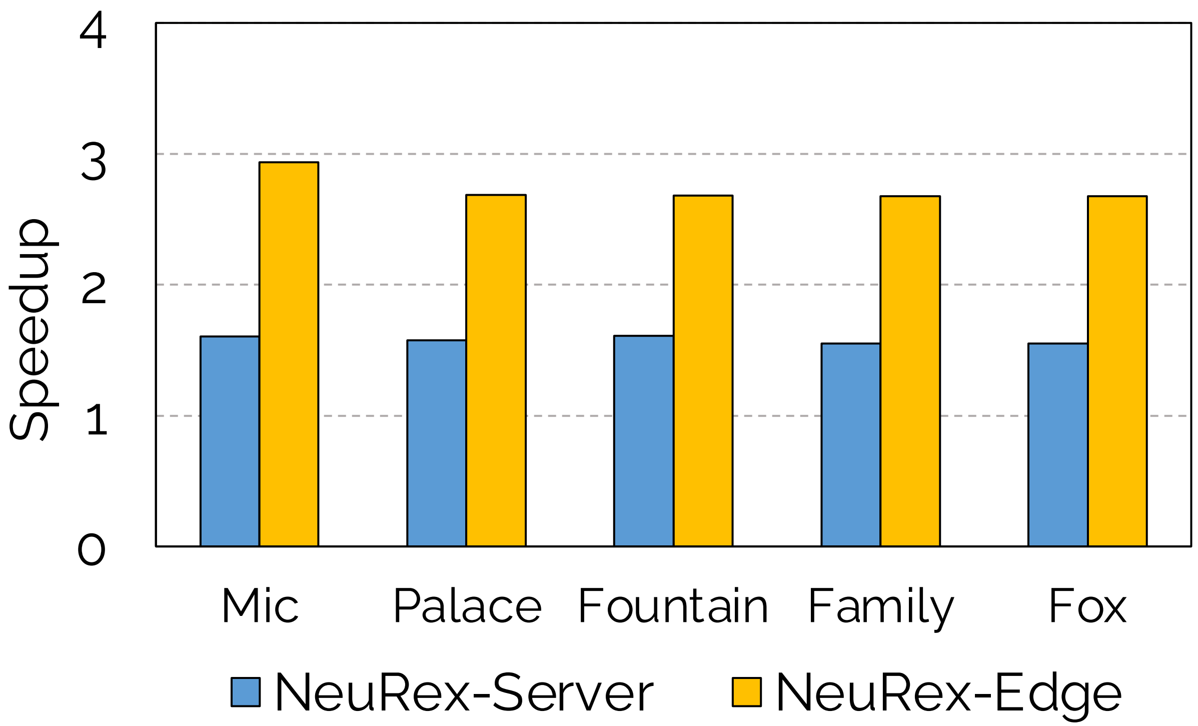
Higher speedup of **NeuRex-Edge**

→ NeuRex enables the performance portability!

Performance: NeRF



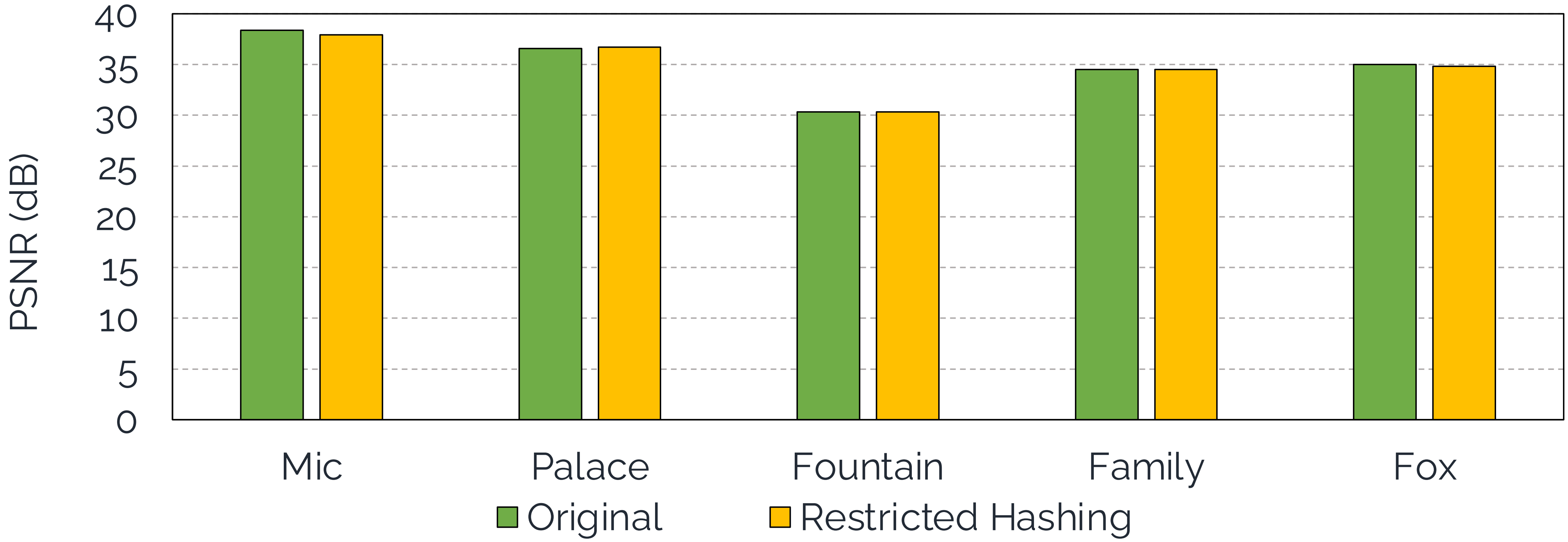
ENC



MLP

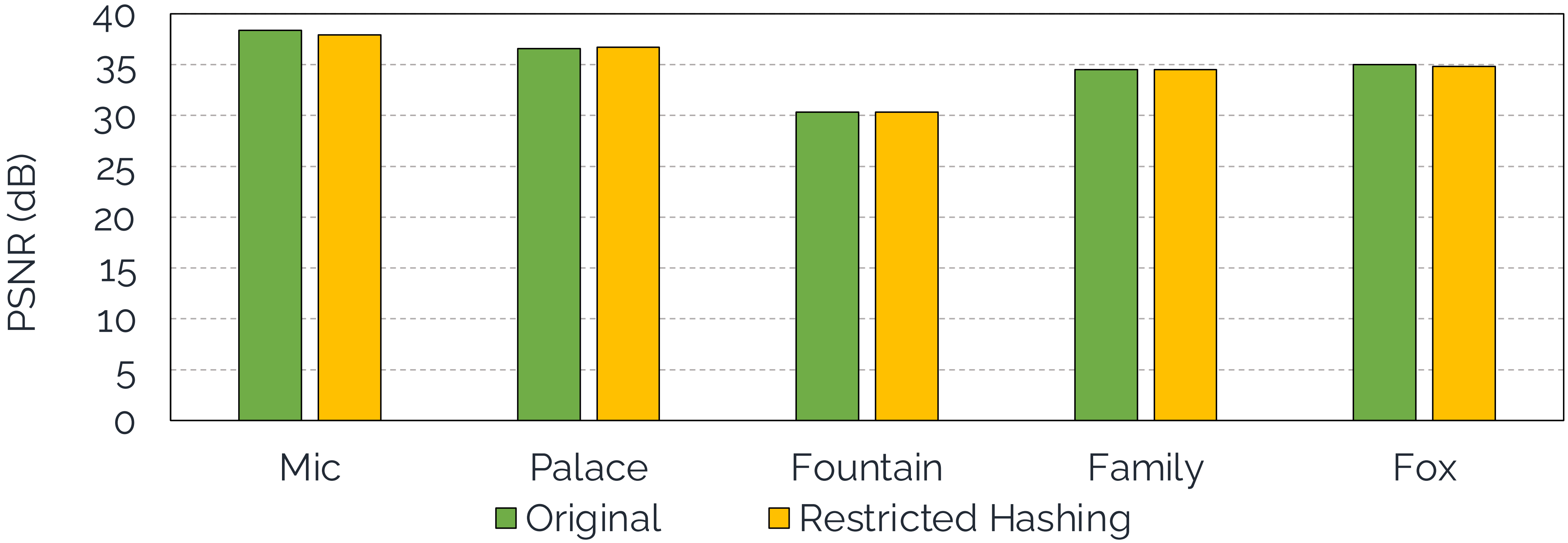
Rendering Quality

Rendering Quality



Rendering Quality

* **Higher is better**



Rendering Quality

Reference

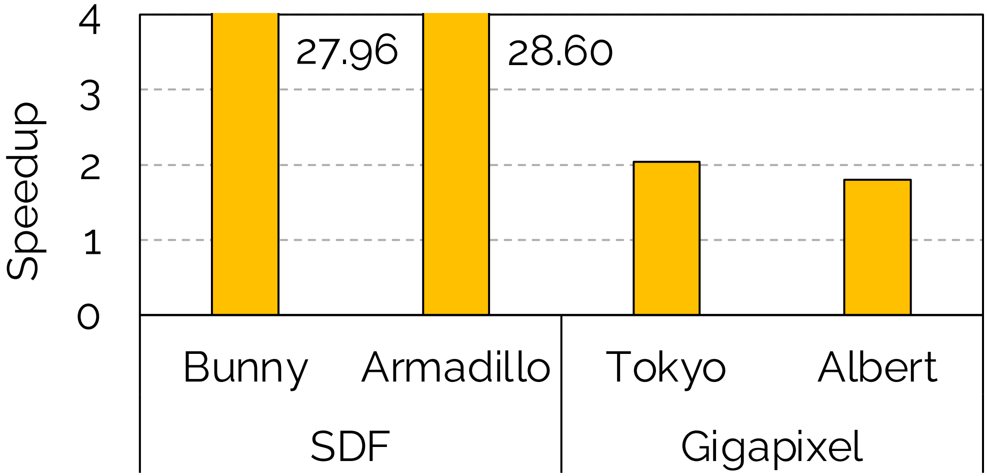
Original Algorithm
(36.63dB)

Restricted Hashing
(36.66dB)

For some scenes, restricted hashing shows
even better rendering quality than the original! ☺

More Details in Our Paper

- Neural Rendering Tasks Beyond NeRFs
  - Signed Distance Functions (SDF)
  - 2D Image Approximation (Gigapixel)



- Source of Performance Gains
- Sensitivity Study
- Area and Energy Efficiency
- Others…

Conclusion

Problem

- Multi-resolution hash encoding is a primary bottleneck in neural rendering with several limitations

Solution: **NeuRex,** an efficient neural rendering accelerator

- *Restricted Hashing* enables performance portability and maximizes resource utilization
- *Neural Graphics Engine* with two specialized on-chip memories

Result

- **NeuRex** achieves up to a 9.88x speedup over the GPU with a substantially small area overhead! ☺

Thank You!