

03

Intro of Semantic IDs

Language Tokenization

Human-readable Data 🧑 🧑

Will be the 46th film overall in the

Machine-readable Data 🤖



Tokenize

8743, 307, 262, 6337, 400, 2646, 4045, 287, 262



9923, 21932, 1512, 11950



Detokenize

Marvel Cinematic Universe

Language Tokenization

Subword Tokenization

e.g., BPE (Byte-Pair Encoding) in gpt-5

Machine-readable Data

8743, 307, 262, 6337, 400, 2646, 4045, 287, 262

Will _be _the _46 th _film _overall _in _the



9923, 21932, 1512, 11950

_Marvel _Cinem atic _Universe

Language Tokenization

Subword Tokenization

e.g., BPE (Byte-Pair Encoding) in gpt-5

Machine-readable Data

8743, 307, 262, 6337, 400, 2646, 4045, 287, 262

Will _be _the _46 th _film _overall _in _the



9923, 21932, 1512, 11950

_Marvel _Cinem atic _Universe

Q1: Why not use **one token per word**?

Language Tokenization

Subword Tokenization

e.g., BPE (Byte-Pair Encoding) in gpt-5

Machine-readable Data

8743, 307, 262, 6337, 400, 2646, 4045, 287, 262  9923, 21932, 1512, 11950
Will _be _the _46 th _film _overall _in _the _Marvel _Cinem atic _Universe

Q1: Why not use **one token per word**?

Large token vocabulary; Not robust enough;

Language Tokenization

Subword Tokenization

e.g., BPE (Byte-Pair Encoding) in gpt-5

Machine-readable Data

8743, 307, 262, 6337, 400, 2646, 4045, 287, 262  9923, 21932, 1512, 11950
Will _be _the _46 th _film _overall _in _the _Marvel _Cinem atic _Universe

Q2: Why not use **one token per character**?

Language Tokenization

Subword Tokenization

e.g., BPE (Byte-Pair Encoding) in gpt-5

Machine-readable Data

8743, 307, 262, 6337, 400, 2646, 4045, 287, 262  9923, 21932, 1512, 11950
Will _be _the _46 th _film _overall _in _the _Marvel _Cinem atic _Universe

Q2: Why not use **one token per character**?

Too many tokens per sentence (long context windows);
Difficult to model;

Action Tokenization

Human-readable Data  



Machine-readable Data 



Tokenize

i3487, i124, i19240, i772

One token per action? 

Action Tokenization

Human-readable Data  



Machine-readable Data 



Tokenize

Premium Men's Short Sleeve Athletic Training T-Shirt Made of Lightweight Breathable Fabric, Ideal for Running, Gym Workouts, and Casual Sportswear in All Seasons; High-Performance Breathable Cotton Crew Socks for Men with Arch Support, Cushioned Heel and Toe, and Moisture Control, Perfect for Sports, Walking, and Everyday Comfort; Men's Loose-Fit Basketball Shorts with Elastic Drawstring Waistband, Quick-Dry Mesh Fabric, and Printed Number 11 for Professional and Recreational Play; Official Size 7 Composite Leather Basketball Designed for Indoor and Outdoor Use, Deep Channel Design for Enhanced Grip and Ball Control, Ideal for Training and Competitive Matches;

Text description of each action? 

Action Tokenization

Can we find a sweet spot between

One token per action

i3487, i124, i19240, i772

Text description

Premium Men's Short Sleeve Athletic Training T-Shirt
Made of Lightweight Breathable Fabric, Ideal for
Running, Gym Workouts, and Casual Sportswear in All
Seasons; High-Performance Breathable Cotton Crew
Socks for Men with Arch Support, Cushioned Heel and
Toe, and Moisture Control, Perfect for Sports, Walking,
and Everyday Comfort; Men's Loose-Fit Basketball
Shorts with Elastic Drawstring Waistband, Quick-Dry
Mesh Fabric, and Printed Number 11 for Professional
and Recreational Play; Official Size 7 Composite
Leather Basketball Designed for Indoor and Outdoor
Use, Deep Channel Design for Enhanced Grip and Ball
Control, Ideal for Training and Competitive Matches;

Semantic IDs

(also called: SemID or SID)

A few tokens that jointly index one item.

t3, t321, t643, t1011



Semantic IDs

(also called: SemID or SID)

A few tokens that jointly index one item.

t3, t321, t643, t1011

{t257, t258, ..., t320, **t321**, t322, ..., t511, t512}

Each token from a **vocabulary shared by all items**

Semantic IDs

(also called: SemID or SID)

A few tokens that jointly index one item.

t3, t321, t643, t1011

Can index maximally $256^4 \approx 4.3 \times 10^9$ items with 1024 tokens

(4 tokens per item, each from a vocabulary of 256)

Action Tokenization

Can we find a sweet spot between

One token per action

i3487, i124, i19240, i772

Text description of each action

Premium Men's Short Sleeve Athletic Training T-Shirt Made of Lightweight Breathable Fabric, Ideal for Running, Gym Workouts, and Casual Sportswear in All Seasons; High-Performance Breathable Cotton Crew Socks for Men with Arch Support, Cushioned Heel and Toe, and Moisture Control, Perfect for Sports, Walking, and Everyday Comfort; Men's Loose-Fit Basketball Shorts with Elastic Drawstring Waistband, Quick-Dry Mesh Fabric, and Printed Number 11 for Professional and Recreational Play; Official Size 7 Composite Leather ...

Action Tokenization

Can we find a sweet spot between

One token per action

i3487, i124, i19240, i772

Semantic IDs

t34, t392, t600, t891, t21,
t502, t592, t1002, t123,
t403, t611, t821, t21,
t502, t711, t1022

Text description of each action

Premium Men's Short Sleeve Athletic Training T-Shirt Made of Lightweight Breathable Fabric, Ideal for Running, Gym Workouts, and Casual Sportswear in All Seasons; High-Performance Breathable Cotton Crew Socks for Men with Arch Support, Cushioned Heel and Toe, and Moisture Control, Perfect for Sports, Walking, and Everyday Comfort; Men's Loose-Fit Basketball Shorts with Elastic Drawstring Waistband, Quick-Dry Mesh Fabric, and Printed Number 11 for Professional and Recreational Play; Official Size 7 Composite Leather ...

Action Tokenization

Can we find a sweet spot between

One token per action

Semantic IDs

Text description of each action

Length

4

16

148

#Vocab

~100k

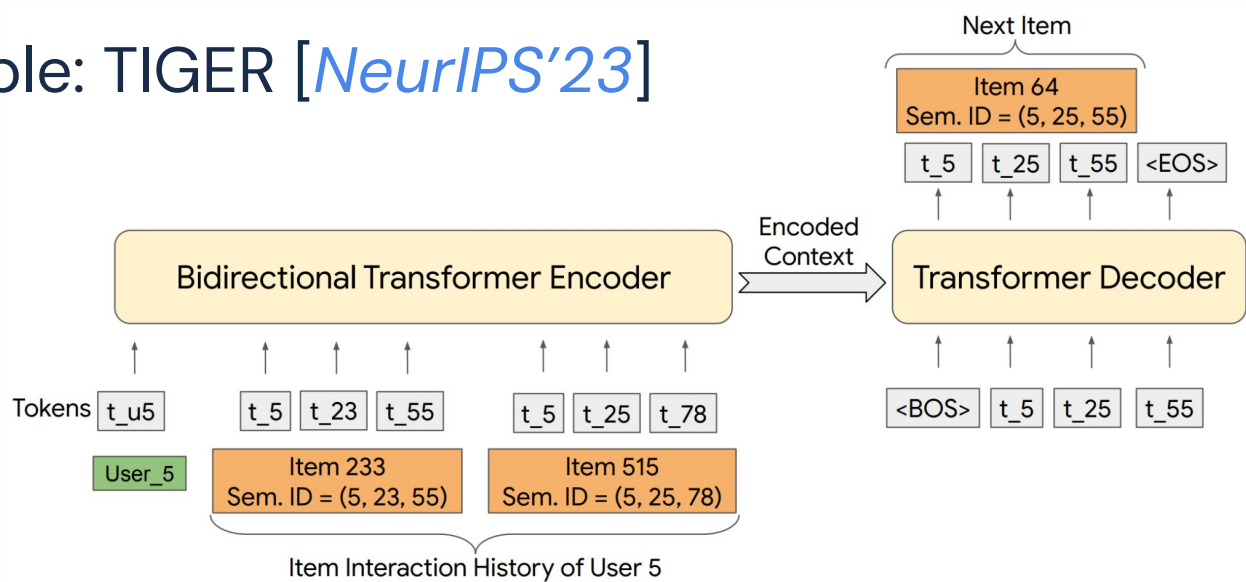
1024

0

Action Tokens

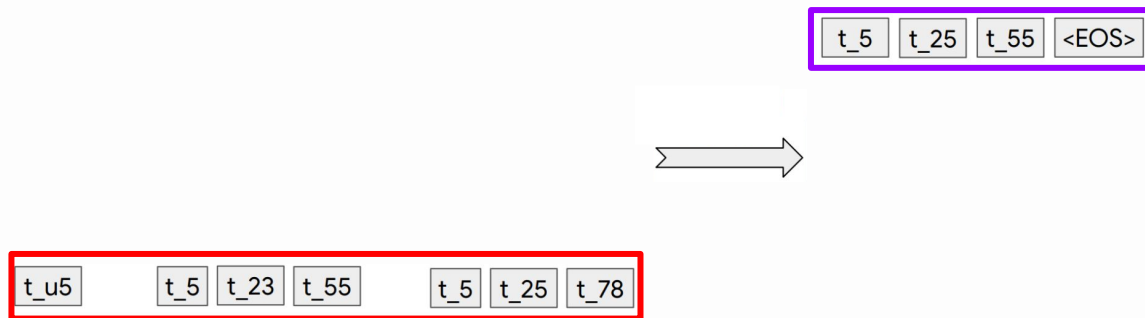
Generative Models based on Semantic IDs

Example: TIGER [*NeurIPS'23*]



Generative Models based on Semantic IDs

Example: TIGER [*NeurIPS'23*]



Recommendation as a **seq**-to-**seq** generation problem

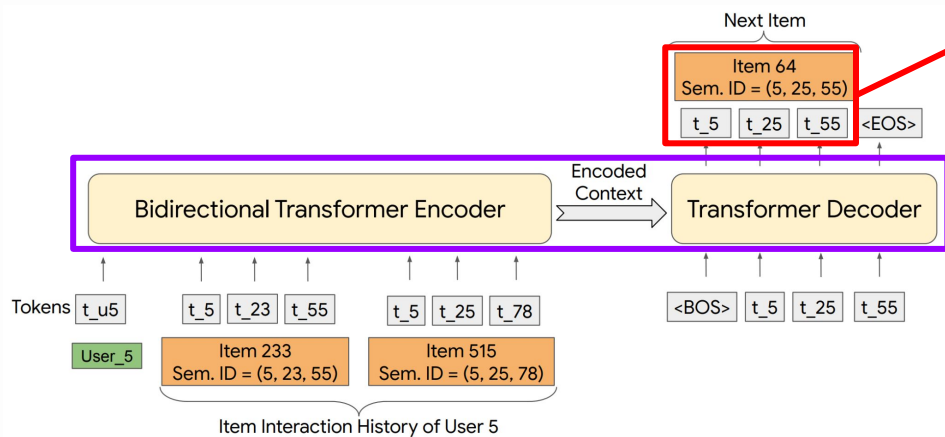
Generative Models based on Semantic IDs

Recommendation as a **seq**-to-**seq** generation problem

Input: user interacted items $\{c_{11'}, c_{12'}, c_{13'}, c_{14'}, c_{21'}, c_{22'}, \dots\}$

Output: next item $\{c_{t1'}, c_{t2'}, c_{t3'}, c_{t4'}\}$

SemID-based Generative Recommendation



Part 1:

How to construct **SIDs**

Part 2:

How to build SID-based
Generative Rec **Models**