



기계는 어떻게 수어를 배울 수 있을까?

How Do Machines Learn to Understand and Produce Sign Language?

Presented by: Xiaohan Ma

Ajou University



The facts about deaf people and sign languages

5%



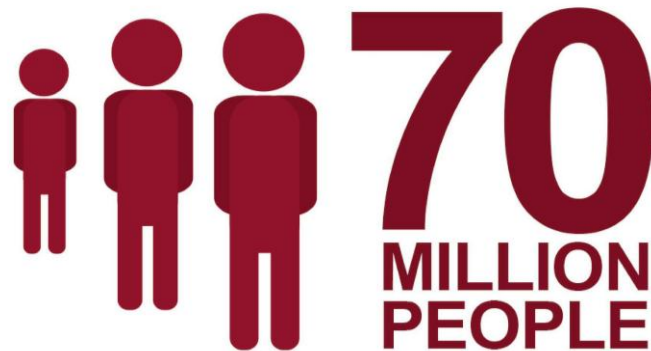
According to the World Health Organization, there are 466 million deaf people in the world (432 million adults and 34 million children)

The World report [1] on hearing envisions

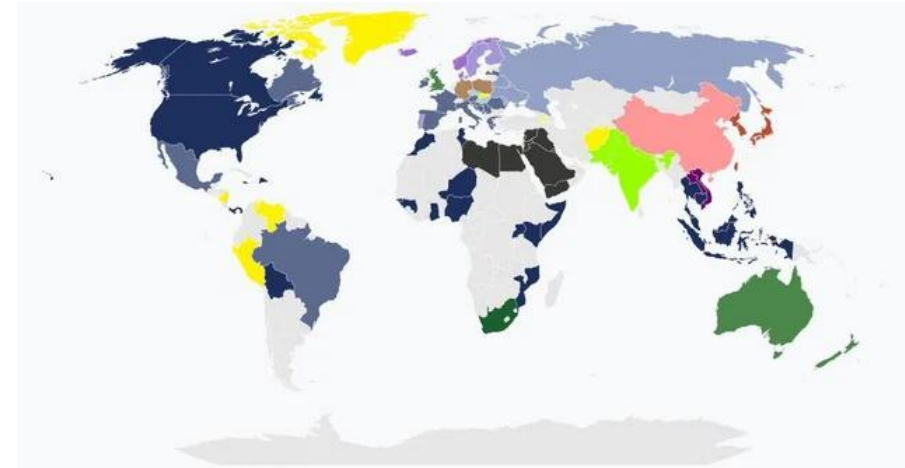
5 INTERESTING FACTS ABOUT SIGN LANGUAGE

SOURCE: LANGUAGES UNLIMITED
WESTON COLLEGE

According to the World Federation of Deaf, about 70 million people in the world use sign language to communicate.



70
MILLION
PEOPLE



The classification of Sign Language families.

	French Sign Language family
	American Sign Language (ASL) cluster, derived from FSL
	Russian Sign Language cluster, derived from FSL
	Czech Sign Language cluster, derived from FSL
	Danish Sign Language family, probably related to either FSL or SSL
	Swedish Sign Language family, probably related to DSL
	German Sign Language family
	Vietnamese sign languages, also some Thai and Lao SLs
	Arab sign-language family
	Indo-Pakistani Sign Language
	Chinese Sign Language (unrelated to Taiwanese Sign Language)
	Japanese Sign Language family (including Taiwanese Sign Language)
	BANZSL family (British, Australian and New Zealand Sign Language)
	South African Sign Language, derived from BANZSL
	Isolated languages
	No data









[1] World Health Organization. 2021. World Report On Hearing - Executive Summary. Technical report.

What is Sign Language and Why It Matters

- Sign language is a complete natural language, with its own **grammar, structure, and rhythm**
- For many Deaf individuals, it is their **primary**—or only—language
- Yet most digital tools support only spoken or written language
- This creates serious barriers to **access, communication,**



YTN News

 Spoken English:	 Sign Language:
✓  Linear order	✓  Spatial layout
✓  Uses tenses	✓  Uses space to show time
✓  Words for pronouns	✓  Points to referents

The Challenge: Machines Can Mimic Signs, But Don't Understand Them

- Sign language is **visual, spatial, and expressive**— hard for AI to learn
- It uses **movement, facial expressions, and 3D space** to convey meaning
- Machines can **track and mimic** sign motions
- But they still **don't understand the language** or its meaning
- This is a long-standing challenge in sign language
- Our work does **not solve** this — but takes a step by helping machines **generate more natural signing**

● Text: 1D, linear input

□ sign language: 3D, visual-spatial

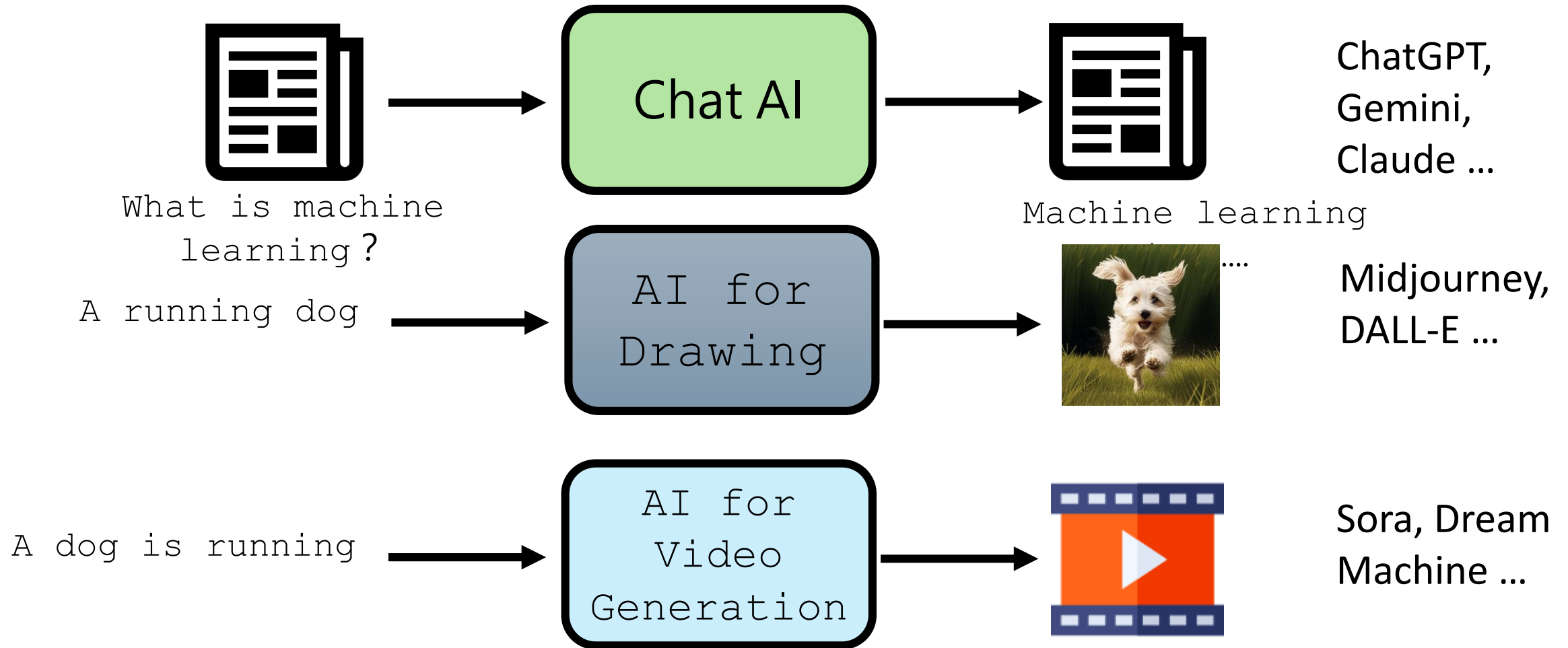
말을 잘못했어요.

(말하다 잘못)



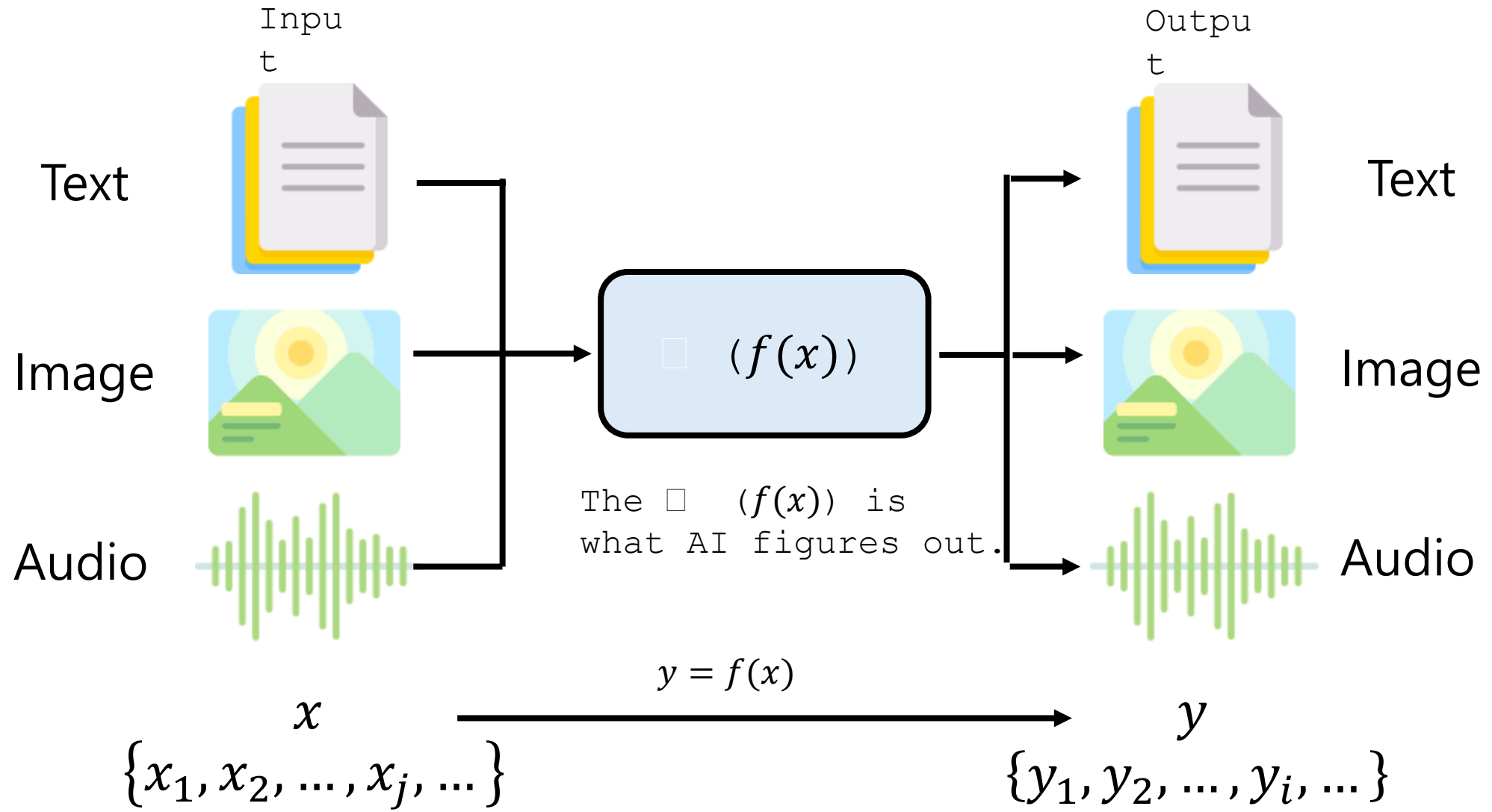
From Text to Vision: What AI Can Do (So Far)

From Text to Image, Music, and Video: What AI Can Already Do



AI \approx Automatically Learning a Function

Instead of us writing the rules, **the machine learns them from data**



How Is Text/Image/Speech Represented for Machines?

- Complex outputs like text, images, and speech can be represented as sequences of **tokens** – small basic units.

$$y = \{y_1, y_2, \dots, y_i, \dots\}$$

Instead of creating everything from scratch, the AI assembles content by **selecting** from known building blocks.



Token

$$y_i = \text{"I"}$$



Token

$$y_i =$$



(image patch)



Token

$$y_i = 0.80$$



(frequency)

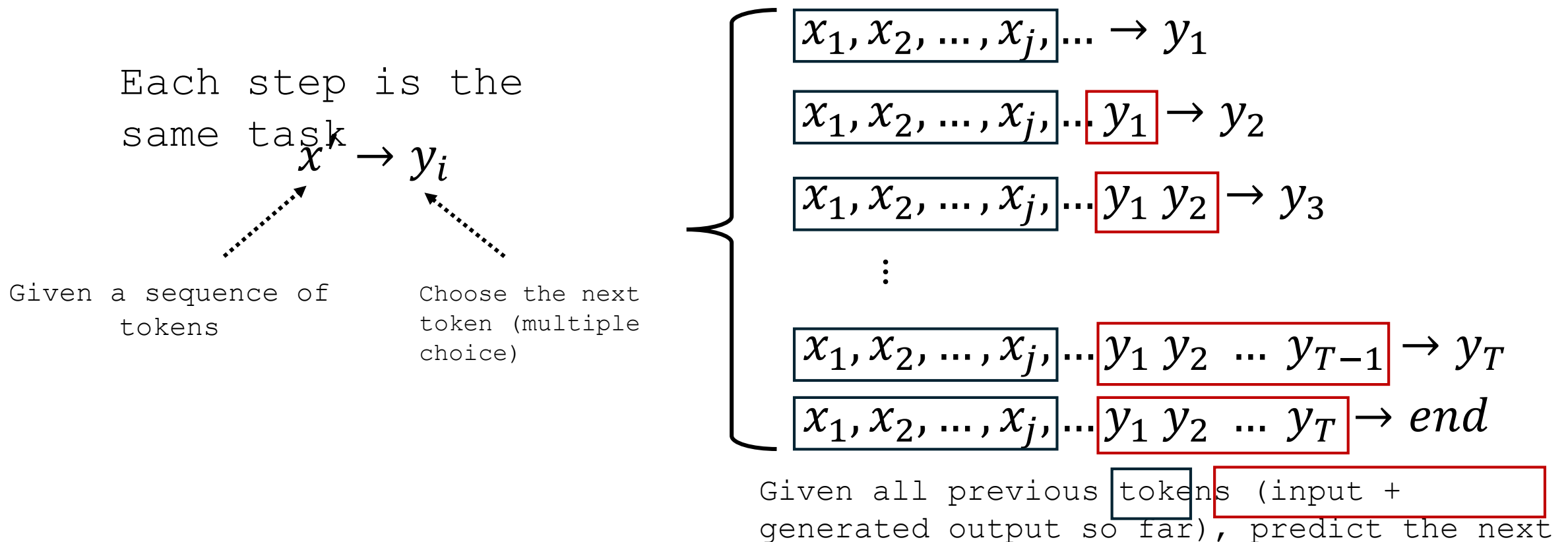
AI selects from a **limited set** of tokens – not infinite guesses.

Token is a basic unit, like a word, image patch, or sound chunk.

How AI Generates Text: One Token at a Time

$$\begin{array}{ccc} x & \xrightarrow{y = f(x)} & y \\ \{x_1, x_2, \dots, x_j, \dots\} & & \{y_1, y_2, \dots, y_i, \dots\} \end{array}$$

Strategy: Generate one y_i at a time in a fixed order

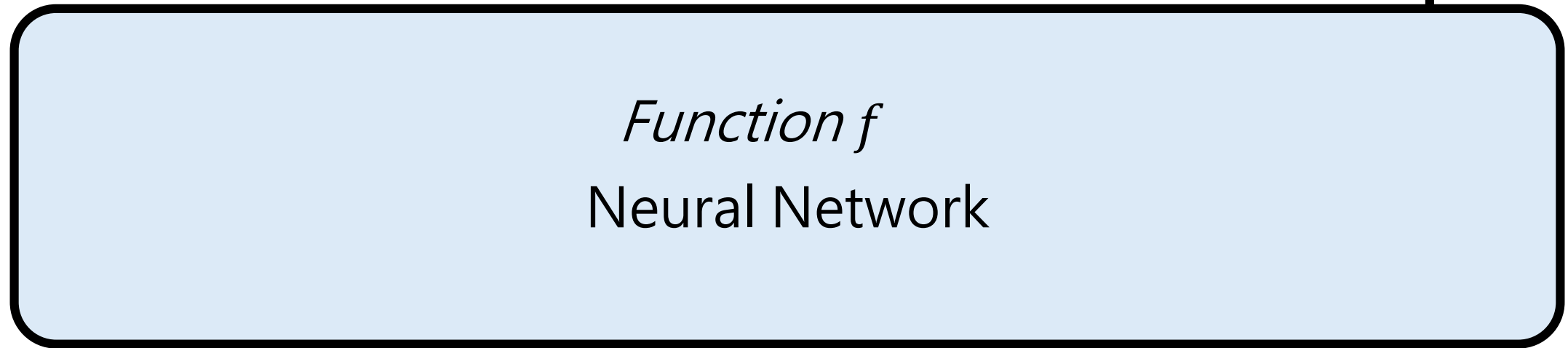
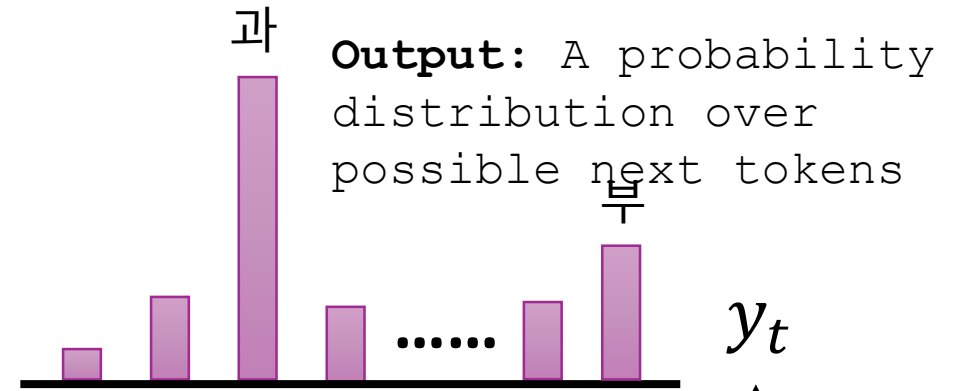


How AI Predicts the Next Token (Using a Function f : A Neural Network)

$$\{y_1, y_2, \dots, y_{t-1}\} \rightarrow y_t$$

인문학

과, 부



y_1

y_2

y_3

.....

Input: Tokens already generated

y_{t-1}

AI Can Read, See, and Hear
— But Can It Sign?

Tokens Can Be More Than Words — So What About Sign?

A screenshot of a YouTube video featuring NVIDIA CEO Jensen Huang. He is standing on a stage, wearing a black leather jacket over a black shirt, and holding a small black object in his right hand and a large NVIDIA graphics card in his left. The background is dark with the NVIDIA logo and the word "NVIDIA" in large letters. The video player interface is visible, including the title "NVIDIA CEO Jensen Huang Keynote at COMPUTEX 2024", a "稍後觀看" (Watch later) button, a "分享" (Share) button, a progress bar at 1:07:06 / 1:49:19, and a "結束全螢幕 (f)" (Exit full screen) button.

NVIDIA CEO Jensen Huang Keynote at COMPUTEX 2024

稍後觀看 分享

Those **tokens** were words, some of the tokens of course could now be images, or charts, or tables, 更多影片 songs ... speech, videos. Those tokens could be anything.

結束全螢幕 (f)

1:07:06 / 1:49:19

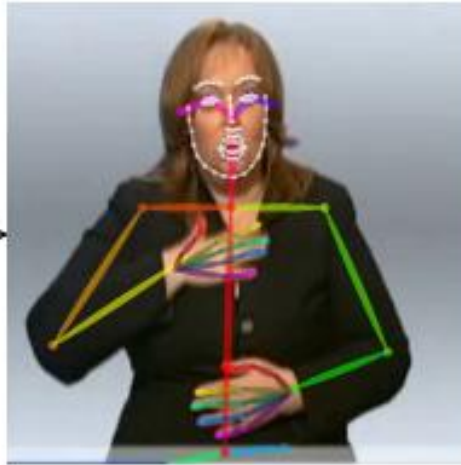
YouTube

How Is Sign Language Represented for Machines?

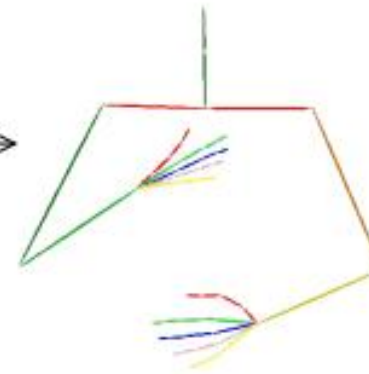
- Sign language is rich and expressive – it uses the **hands**, **face**, and **body** to communicate.
- Using motion capture or pose estimation, each frame is represented by a set of **3D keypoints**.
- To a machine, it becomes a series of **3D numbers** – like sheet music, but for the **whole body**.
- Each frame is like a **token** – but unlike words, it's made of **movement across space**



Input: Raw
video



Pose
estimation
(2D overlay)



3D joint representation
for machine input

120 keypoints per
frame \times 3 (x, y,
z) = 360 values

How AI Generates Sign Language: One Frame at a Time

- It's like building a sentence – but using **body movement** instead of words
- AI generates signing **one frame at a time**
- Each frame encodes a complete pose of the **hands, face, and body**
- The model predicts **what comes next**, using the **input text** and **previous frames**

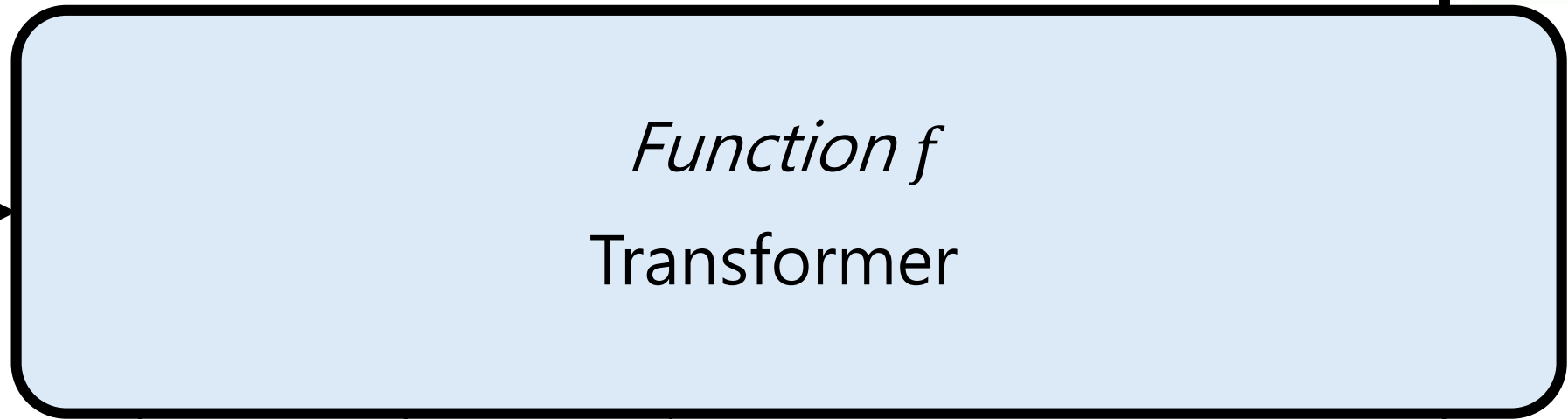
□ Output: One full-body pose (hands + face + body)



🗨️ Input:

서울역 가다 목적
지하철 번호

English Translation:
The subway line
number for the
destination Seoul
station



y_1



y_2



y_3

.....

Previous Frames (as input)



y_{t-1}

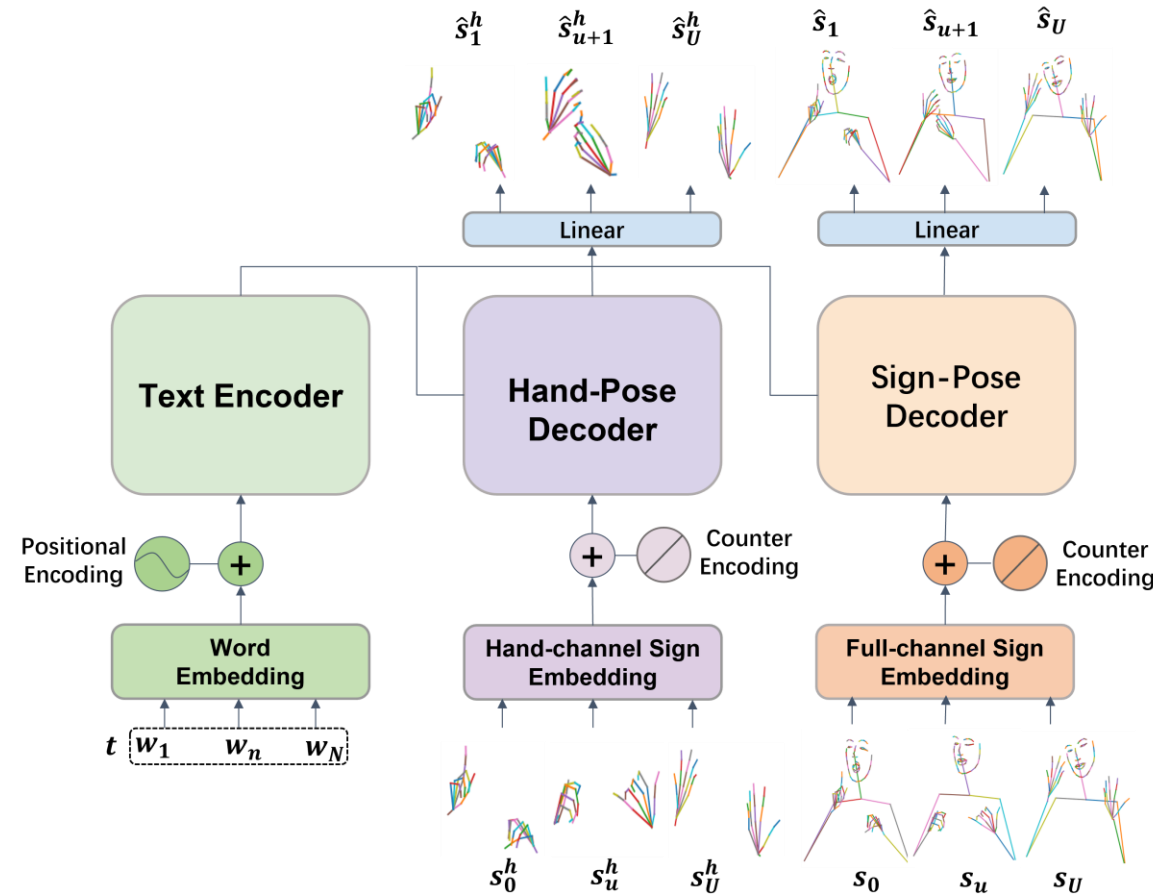


Our Approach: How We Solve It?

Work 1: A Two-Step Way to Teach AI

Signing

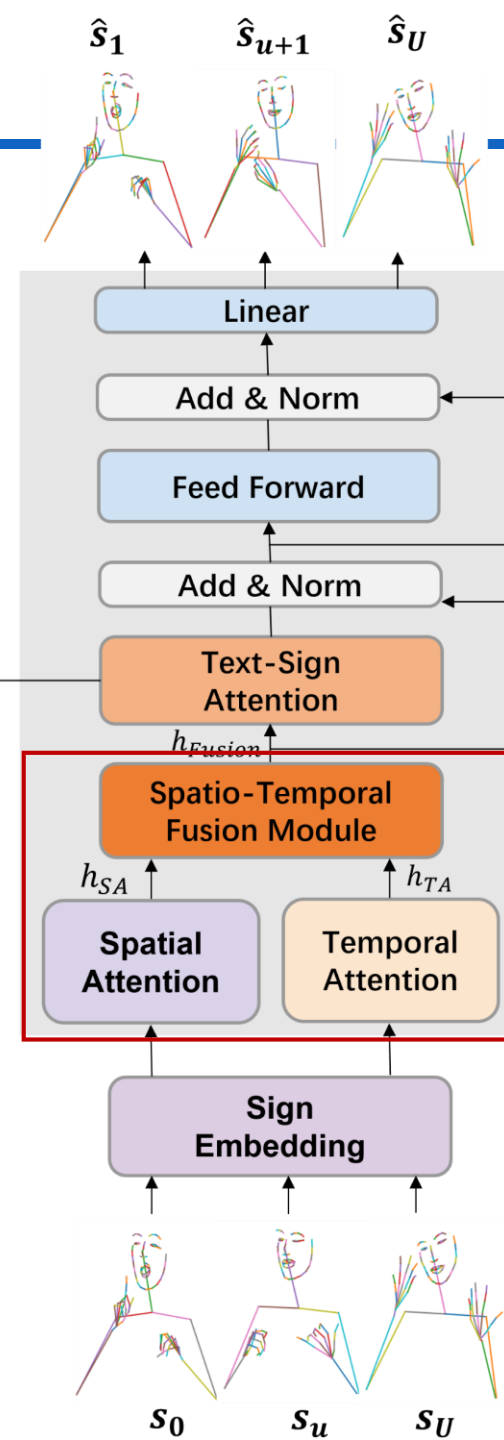
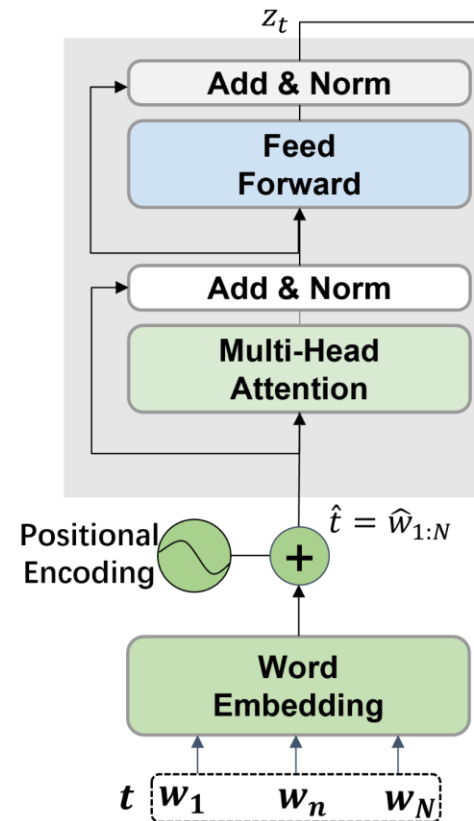
- We first teach the model to sign using **only the hand movements**
- Then we let it expand to **full-body motion**
- This step-by-step method helps the model produce **clearer and more natural** signs



Our Approach: How We Solve It?

Work 2: Teaching AI to Coordinate All Channels

- The model learns to coordinate hands, face, and body together
- It considers both space and time for smooth and natural motion.
- This helps the AI produce more expressive and complete signing



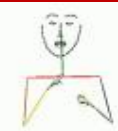
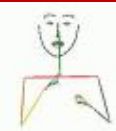
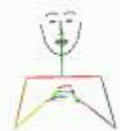
Can AI Really Sign? Let's See the Results

서울역 가다 목적 지하철 번호

English Translation: The subway line number for the destination Seoul Station.

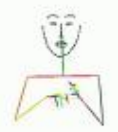
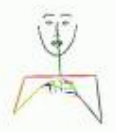
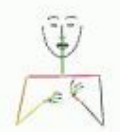
Input

Previous
work



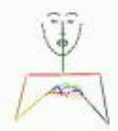
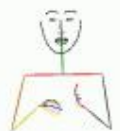
✗ Rigid, hand shape unclear

Work 1



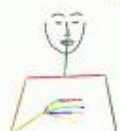
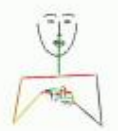
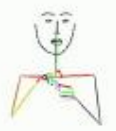
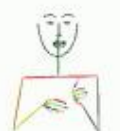
✓ Smoother, hand shape better

Work 2



✓✓ More expressive and natural – hand shapes clearer, timing smoother

Ground
Truth



Original
Image



★ Human signer reference

Frame #

1

2

3

4

5

6

7

What This Work Shows — and What Comes Next

◆ What This Work Shows

- ◆ AI can start generating full-body sign language, one frame at a time
- ◆ Two-step training improves clarity and fluency
- ◆ Coordination across hands, face, and body adds expressiveness
- ◆ Moving toward more natural, full-channel signing

◆ What Comes Next

- ◆ Explore SMPL-X for richer 3D body modeling
- ◆ Use diffusion models to improve video smoothness and realism
- ◆ Expand to diverse signing styles and languages

SMPL-X





Thank you

감사합니다