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

The World report [1] on hearing envisions

[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

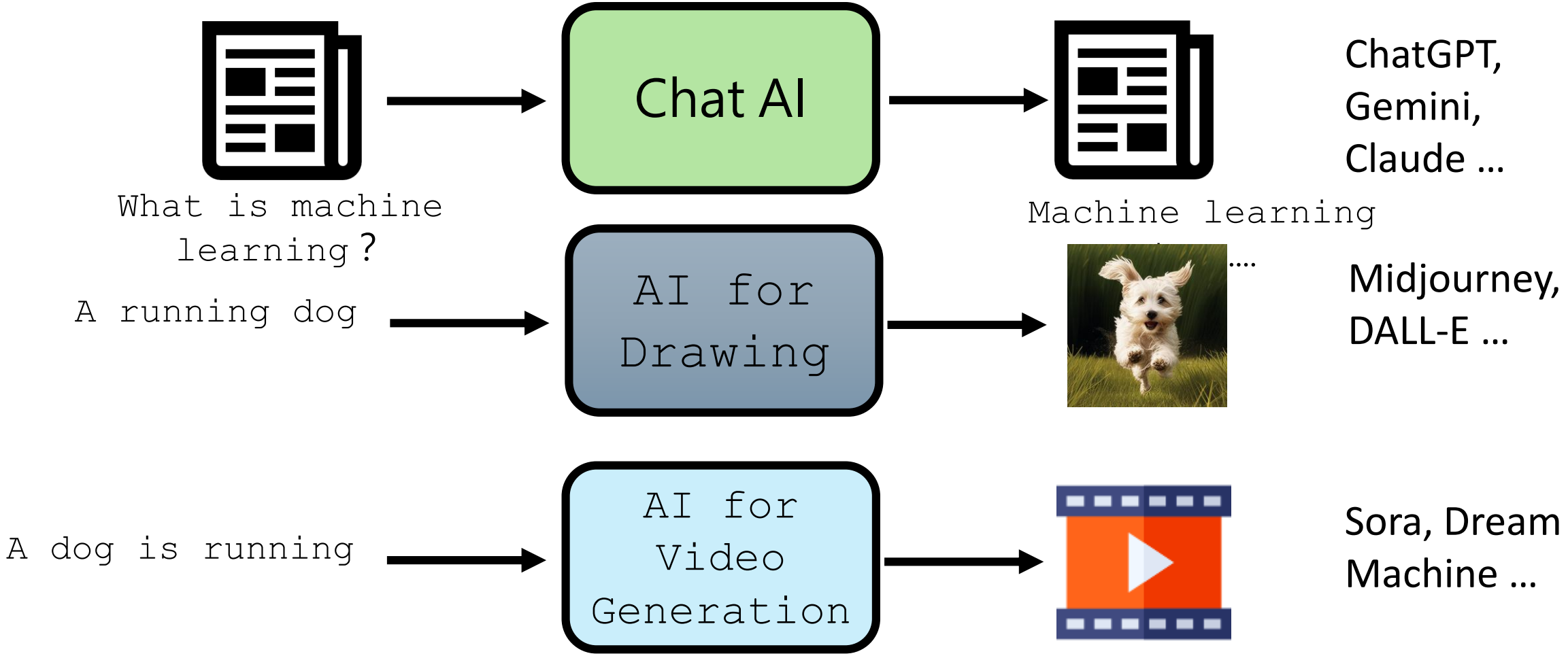
sign language: 3D, visual-spatial

말을 잘못했어요.

(말하다 잘못)

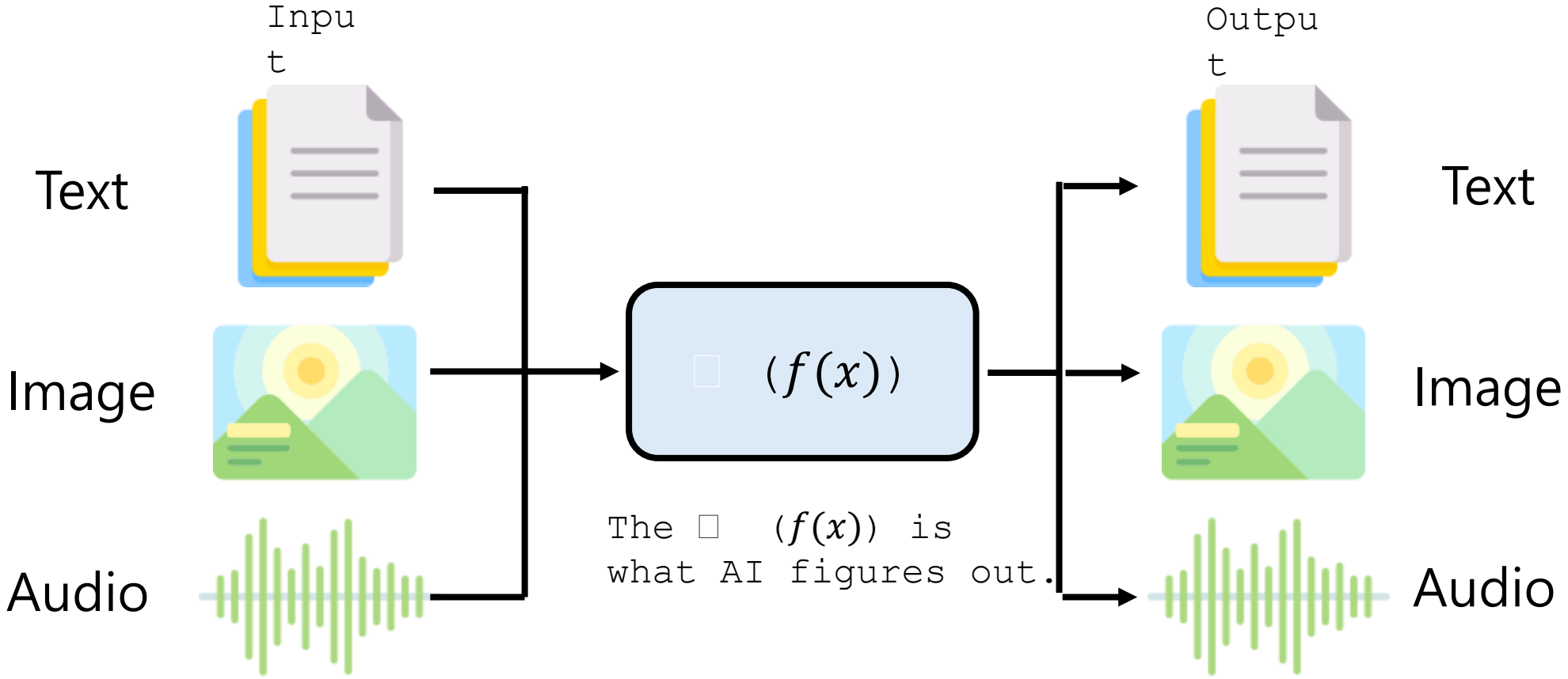
From Text to Vision: What AI Can Do (So Far)

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



AI ≈ Automatically Learning a Function

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



$$x \xrightarrow{y = f(x)} y$$

$$\{x_1, x_2, \dots, x_j, \dots\} \qquad \{y_1, y_2, \dots, y_i, \dots\}$$

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.**

$y =$ 

$y =$ 

$y =$ 

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

Token

$y_i =$ "I"

Token

$y_i =$ 

(image patch)

Token

$y_i = 0.80$

(frequency)

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

How AI Generates Text: One Token at a Time

$$x \xrightarrow{y = f(x)} y$$

$$\{x_1, x_2, \dots, x_j, \dots\} \qquad \{y_1, y_2, \dots, y_i, \dots\}$$

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

Each step is the same task

$$x \rightarrow y_i$$

Given a sequence of tokens

Choose the next token (multiple choice)

$$\boxed{x_1, x_2, \dots, x_j,} \dots \rightarrow y_1$$

$$\boxed{x_1, x_2, \dots, x_j,} \dots \boxed{y_1} \rightarrow y_2$$

$$\boxed{x_1, x_2, \dots, x_j,} \dots \boxed{y_1\ y_2} \rightarrow y_3$$

$$\vdots$$

$$\boxed{x_1, x_2, \dots, x_j,} \dots \boxed{y_1\ y_2\ \dots\ y_{T-1}} \rightarrow y_T$$

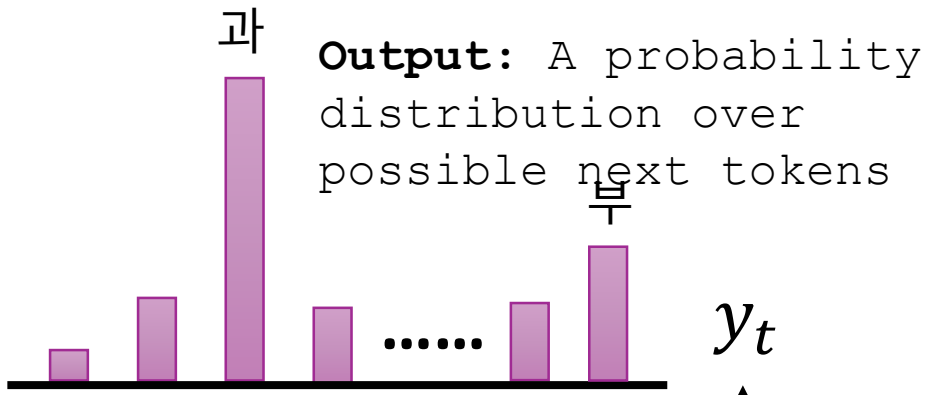
$$\boxed{x_1, x_2, \dots, x_j,} \dots \boxed{y_1\ y_2\ \dots\ y_T} \rightarrow end$$

Given all previous $\boxed{\text{tokens}}$ $\boxed{\text{(input + generated output so far)}}$, predict the next

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 .....



Output: A probability distribution over possible next tokens

y_t

Function f

Neural Network

y_1 y_2 y_3 ..... y_{t-1}

Input: Tokens already generated

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

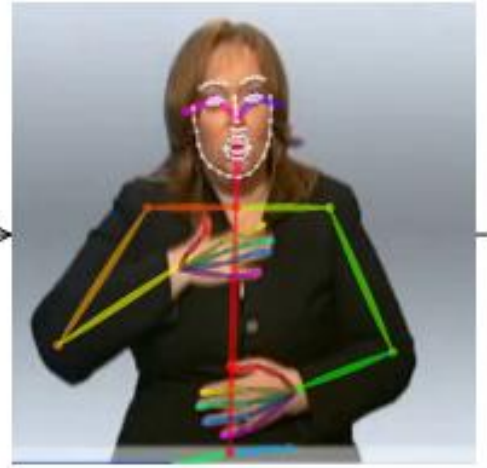
Tokens Can Be More Than Words — So What About Sign?

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

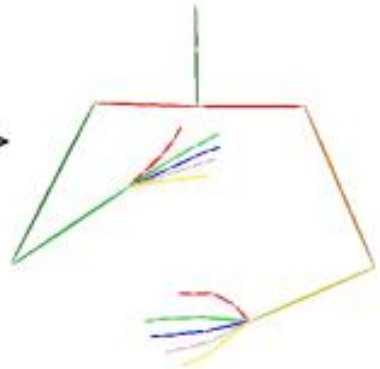
How Is Sign Language Represented for Machines?

- 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)

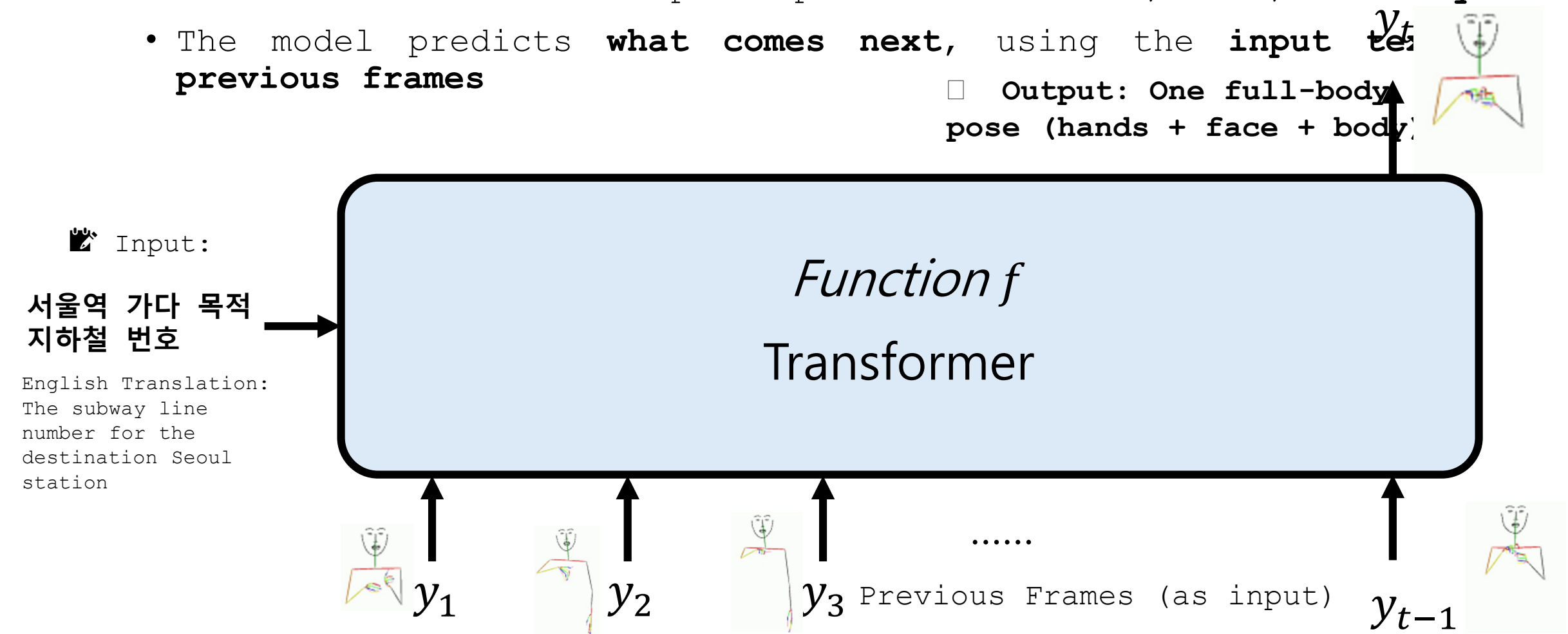


3D joint representation for machine input

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

How AI Generates Sign Language: One Frame at a Time

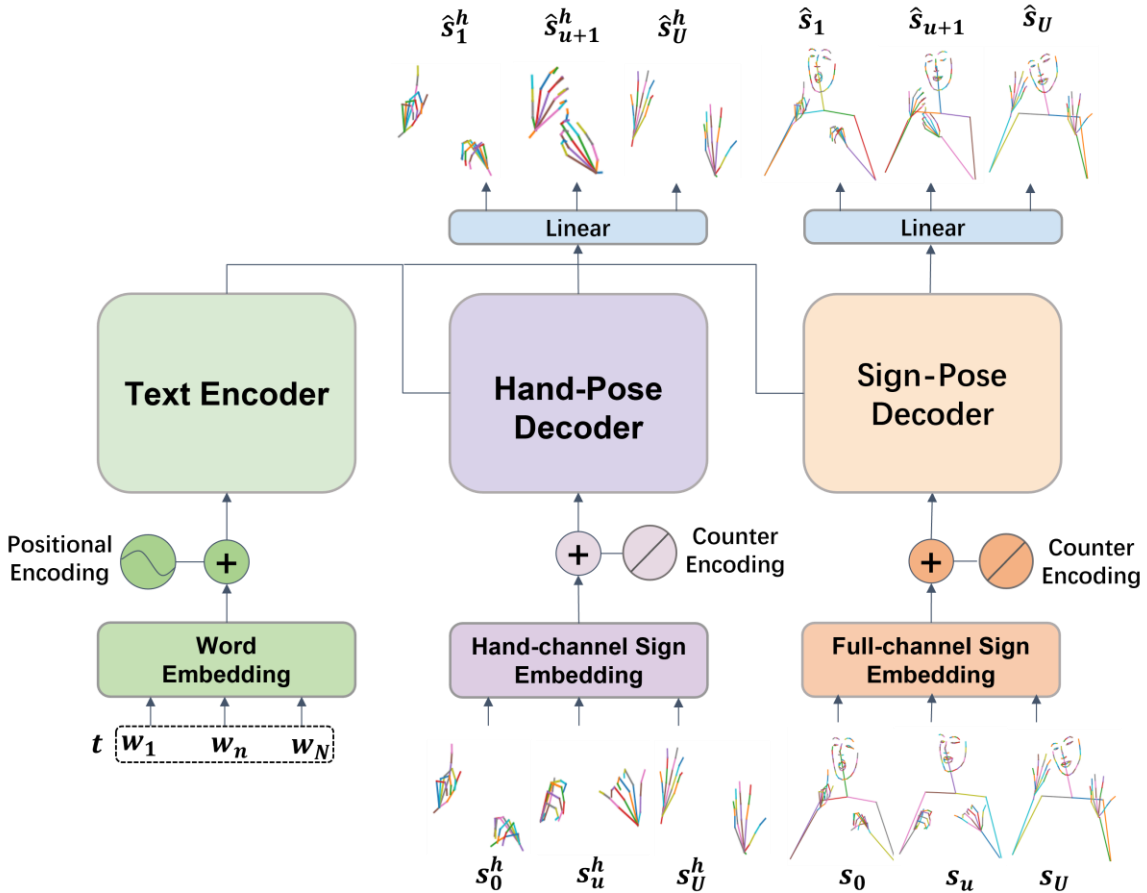
- It's like building a sentence — but using **body movement** instead of words
- AI generates signing **one frame at a time**
- Each frame encodes a complete pose of the **hands**, **face**, and **body**
- The model predicts **what comes next**, using the **input text** + **previous frames**



Our Approach: How We Solve It?

Work 1: A Two-Step Way to Teach AI Signing

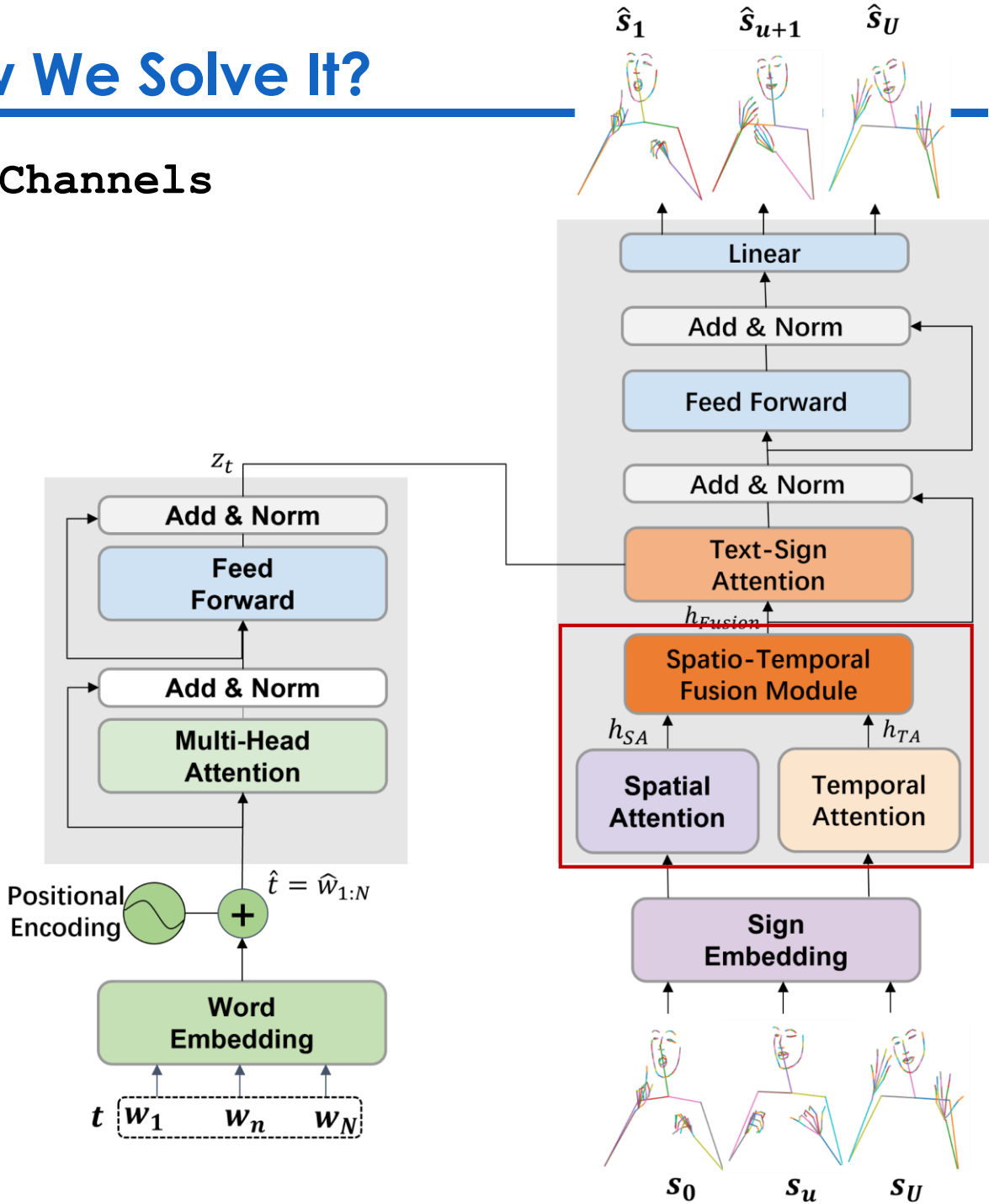
- We first teach the model to sign using **only the hand movements**
- Then we let it expand to **full-body motion**
- This step-by-step method helps the model produce **clearer and more natural** signs



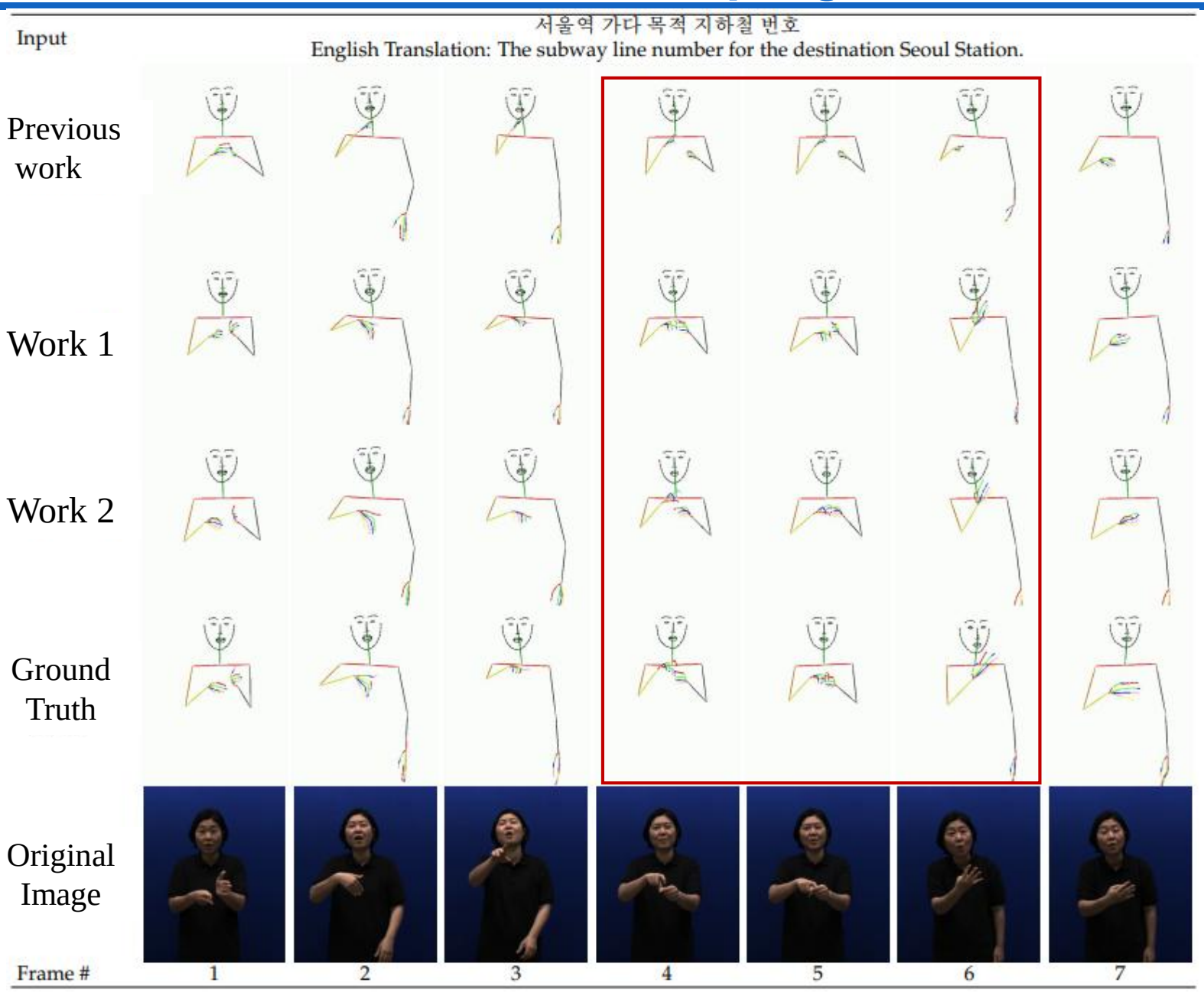
Our Approach: How We Solve It?

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



❌ Rigid, hand shape unclear

✅ Smoother, hand shape better

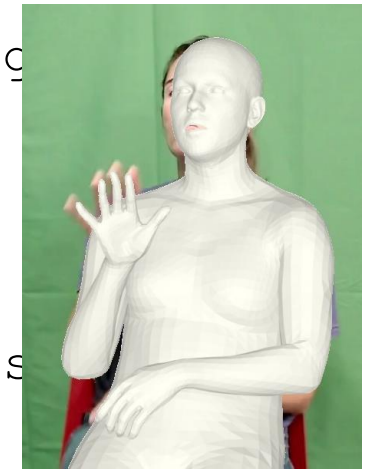
✅✅ More expressive and natural — hand shapes clearer, timing smoother

⭐ Human signer reference

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 realism
  - Expand to diverse signing styles and languages

SMPL-X



Thank you

감사합니다