1st International Workshop on Multimedia Content Generation and Evaluation (McGE), October 29, 2023

Nonword-to-Image Generation Considering Perceptual Association of Phonetically Similar Words

Chihaya Matsuhira, Marc A. Kastner, Takahiro Komamizu, Takatsugu Hirayama, Keisuke Doman, Ichiro Ide

Nagoya University, Japan

Text-to-Image (T2I) generation

- Recent innovation in **Text-to-Image (T2I) generation** models
  - Stable Diffusion[1] is one such example

Example of T2I generation using Stable Diffusion

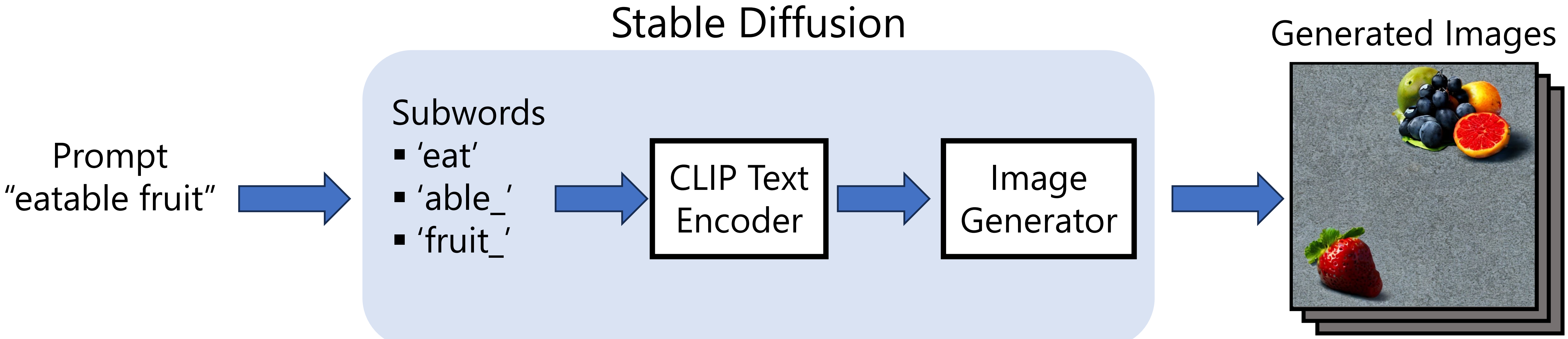
a hamburger floating
in the sky

an astronaut swimming
under the sea

[1] Rombach et al., "High-resolution image synthesis with latent diffusion models", CVPR 2022.

How Stable Diffusion works

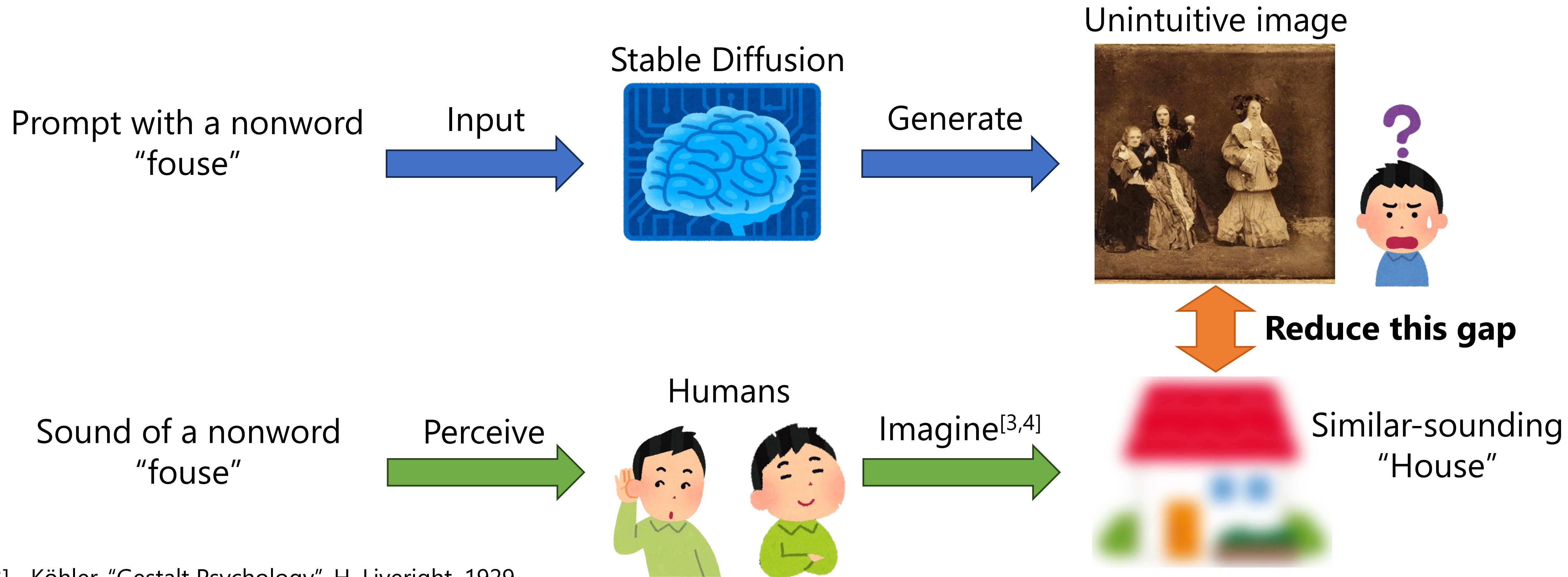
- Stable Diffusion[1]: Open-source text-to-image generation model
  - Generates images from embeddings of the CLIP text encoder
  - CLIP[2]: Vision & language foundation model
    - Consists of text and image encoders co-trained via contrastive learning
    - Subword tokenization: Tokenizes each word in a text into subwords



[2] Radford et al., “Learning transferable visual models from natural language supervision”, ICML 2021.

Problem of T2I Generation Models: Nonword Input

- They generate unintuitive images when input contains **nonwords**
  - Nonwords := “Nonsense words that have no definition within a language”

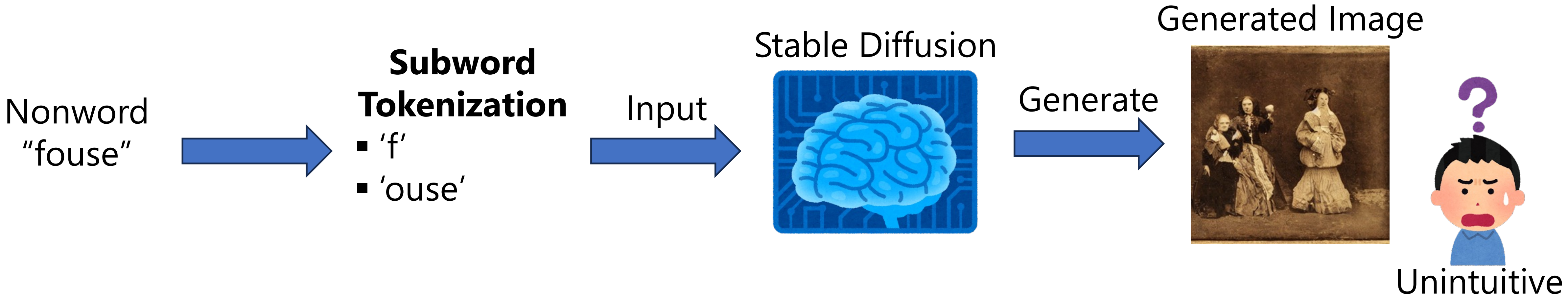


[3] Köhler, “Gestalt Psychology”, H. Liveright, 1929.
[4] Goldinger et al., “Form-based priming in spoken word recognition: The roles of competition and bias”, J. Exp. Psychol. Learn. Mem. Cogn., 1992.

Problem of T2I Generation Models: Tokenization

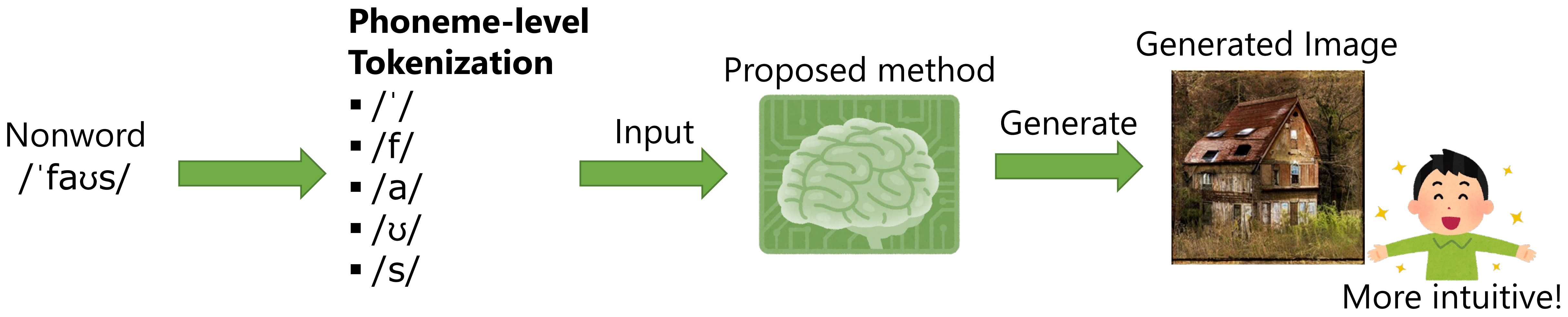
- **Subword tokenization** does not work for nonwords
  - It splits nonwords into **unmeaningful** subwords
    - "fouse" → 'f' + 'ouse' (two subword tokens)
    - Cf. "house" → 'house' (one token)

- Making nonword-to-image generation unintuitive



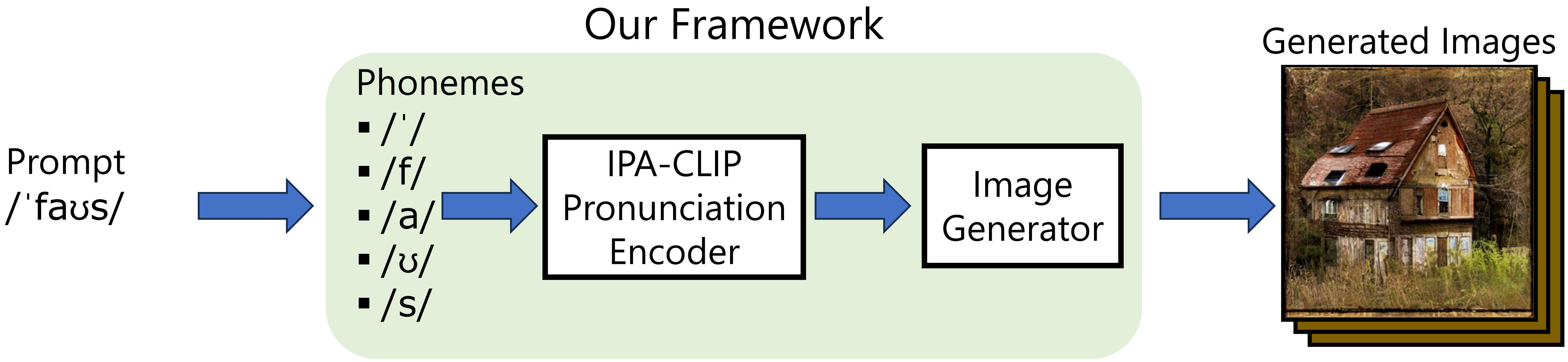
Research Goal

- **More intuitive nonword-to-image generation**
- Approach
  - Replace CLIP text encoder with our new pronunciation encoder
    - ➢ Discard the use of **subword tokenization**
    - ➢ Our **phoneme-level tokenization** considers **phonetic similarity** of an input



Proposed Method: Pronunciation-to-Image Generation

- Our framework consists of two modules:
  - **Pronunciation Encoder**: Pronunciation -> CLIP embedding
  - **Image Generator (Stable Diffusion)**: CLIP embedding* -> Images



IPA-based Phoneme Embedding (1/2)

- IPA: “International Phonetic Alphabet”
- IPA chart[5] is used as a source of phonetic relationships
  - Defines **phonetic properties** of each phoneme/phone in any language
  - Enables computing phonetic similarity

- Compute a **magnitude vector** for each phoneme

IPA Chart for Consonants

	Bilabial		Labio-dental		Dental		Alveolar		Post-alveolar		Palatal		Velar		Glottal	
Nasal		m		ɱ				n				ɲ		ŋ		
Plosive	p	b					t	d			c	ɟ	k	g	ʔ	
Sibilant affricate							ts	dz	tʃ	dʒ	tɕ	dʑ				
Sibilant fricative							s	z	ʃ	ʒ	ɕ	ʑ				
Nonsibilant fricative	ɸ	β	f	v	θ	ð					ç	ʝ	x	ɣ	h	ɦ
Approximant				ʋ				ɹ				j		ɰ		
Lateral approximant								l				ʎ		ʟ		

Consonant /p/
- Unvoiced
- Bilabial
- Plosive

Magnitude Vector

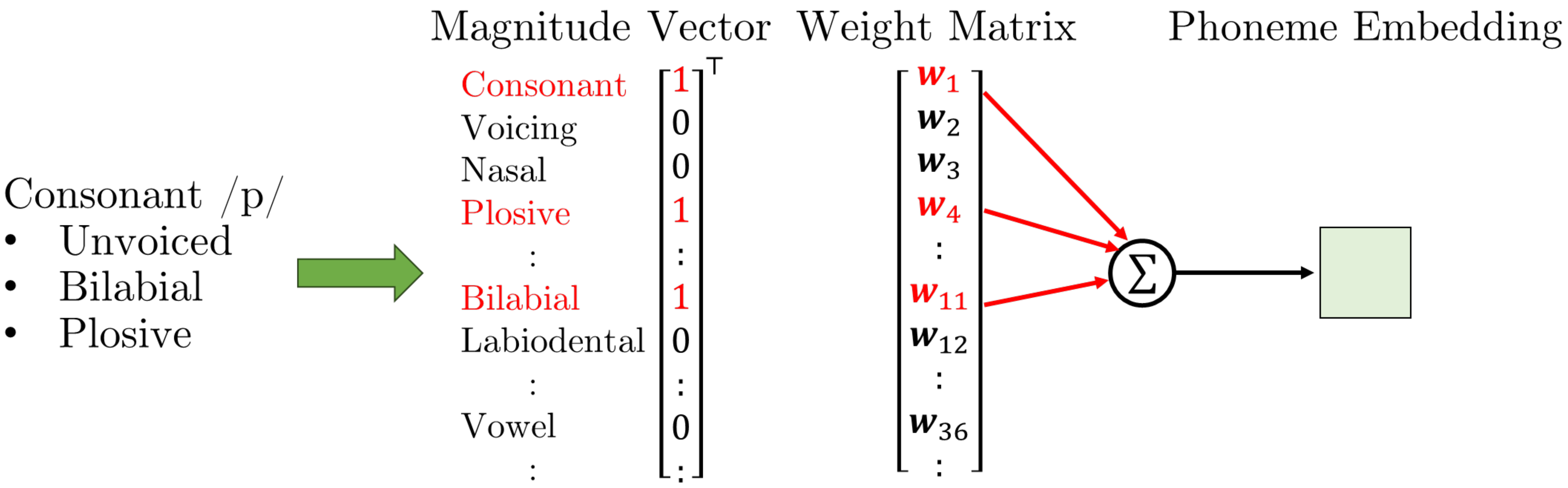
$$\begin{array}{l} \text{Consonant} \\ \text{Voicing} \\ \text{Nasal} \\ \text{Plosive} \\ \vdots \\ \text{Bilabial} \\ \text{Labiodental} \\ \vdots \\ \text{Vowel} \\ \vdots \end{array} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \end{bmatrix}^{T}$$

[5] International Phonetic Association, Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet, Cambridge University Press, 1999.

IPA-based Phoneme Embedding (2/2)

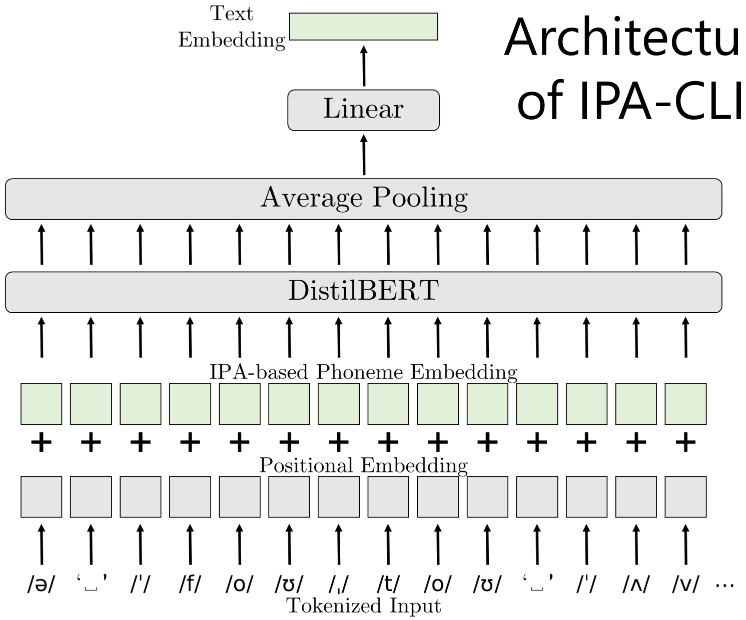
Aim to assign a **phonetically continuous** token for each phoneme

1. Prepare magnitude vector based on phonetic property
2. Multiply it with a trainable weight matrix
3. Obtain a phoneme embedding reflecting the phonetic property

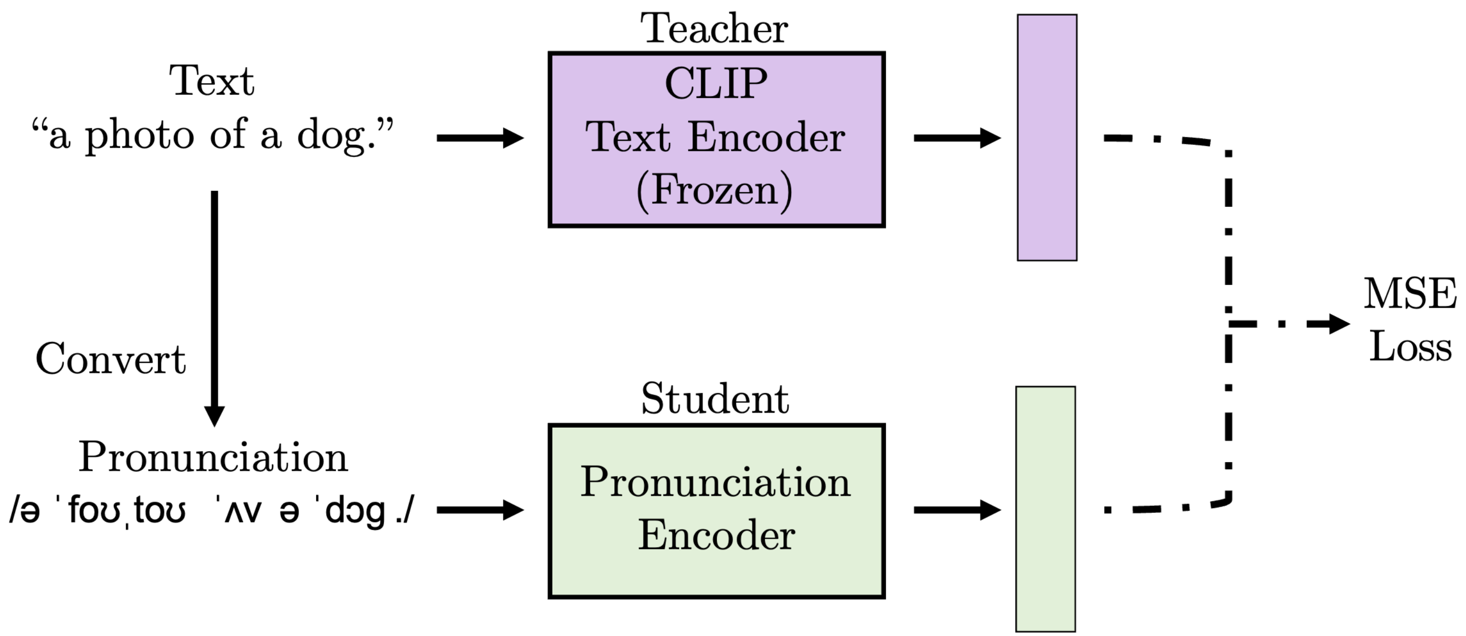


Distillation of CLIP Text Encoder

- Distill the CLIP text encoder with text-pronunciation pairs
  1. Prepare pronunciation for each text in training data[6]
     - ➢ Use existing pronunciation dictionaries
  2. Train a student encoder (**IPA-CLIP**) to output the identical embedding to the teacher encoder with the corresponding pronunciation input



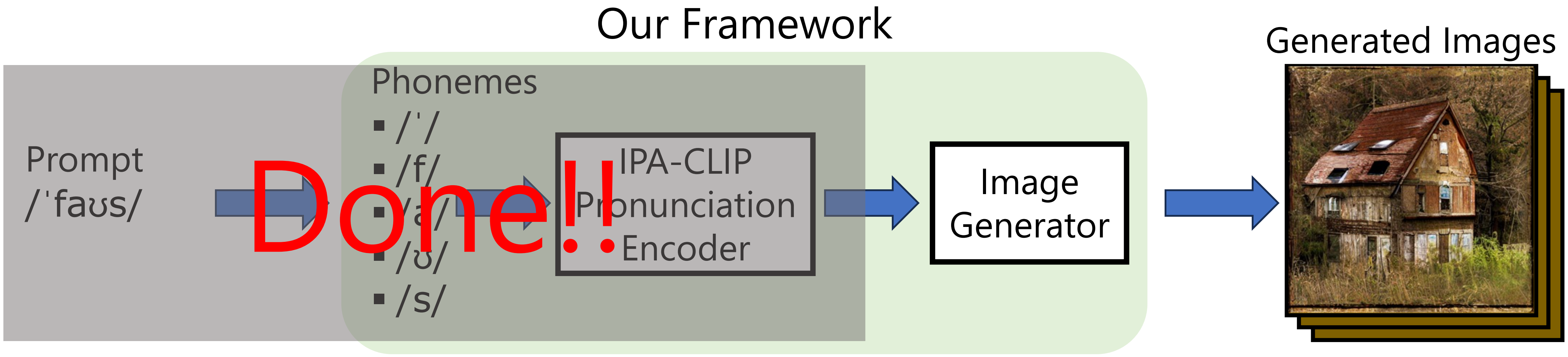
Architecture of IPA-CLIP



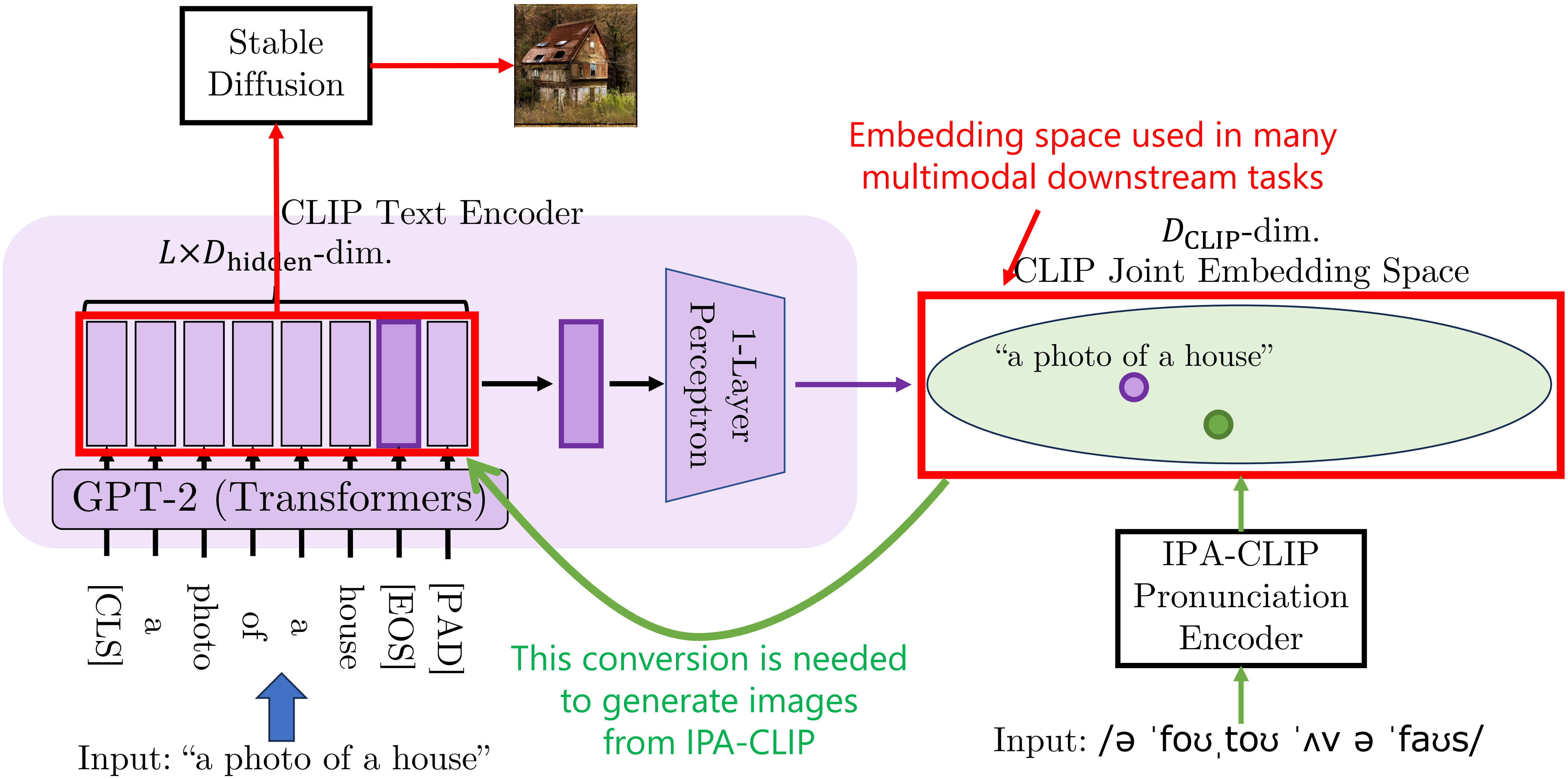
[6] Carlsson et al., "Cross-lingual and multilingual CLIP", LREC 2022.

Proposed Method: Pronunciation-to-Image Generation

- Our framework consists of two modules:
  - **Pronunciation Encoder**: Pronunciation -> CLIP embedding
  - **Image Generator (Stable Diffusion)**: CLIP embedding* -> Images

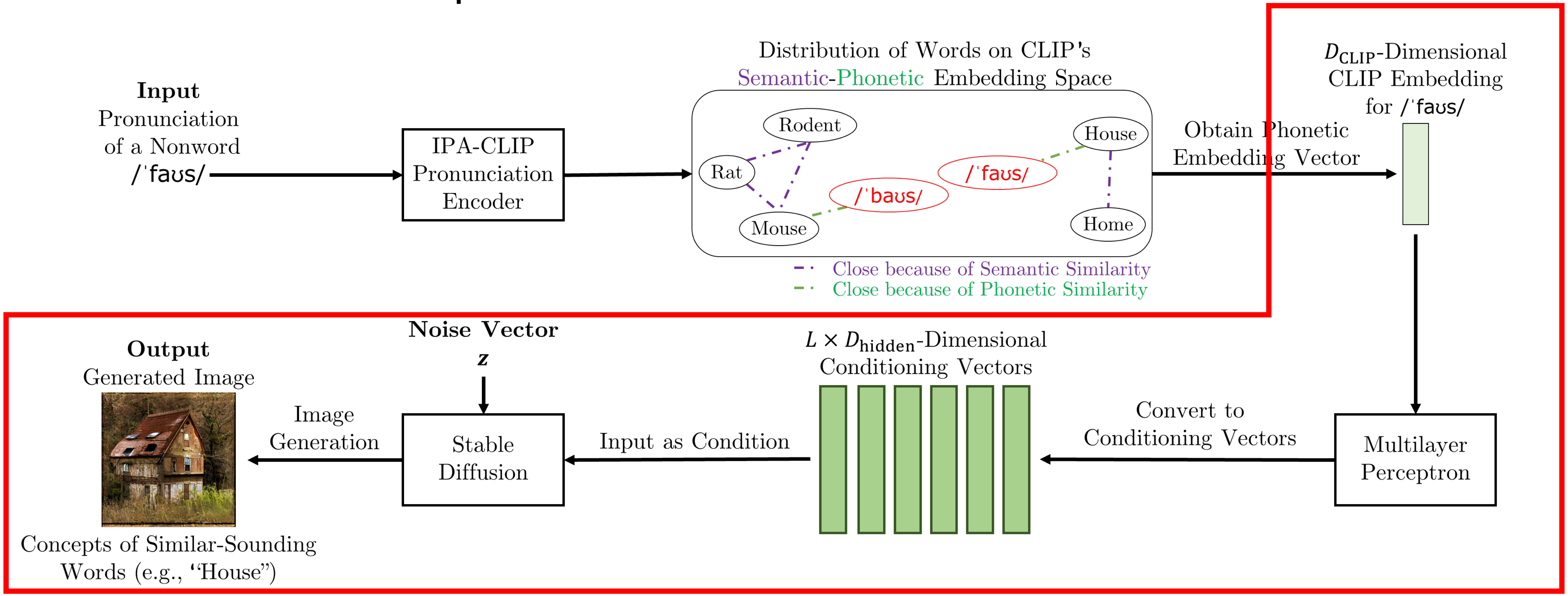


CLIP Text Encoder Explained in Detail



Pronunciation-to-Image Generation

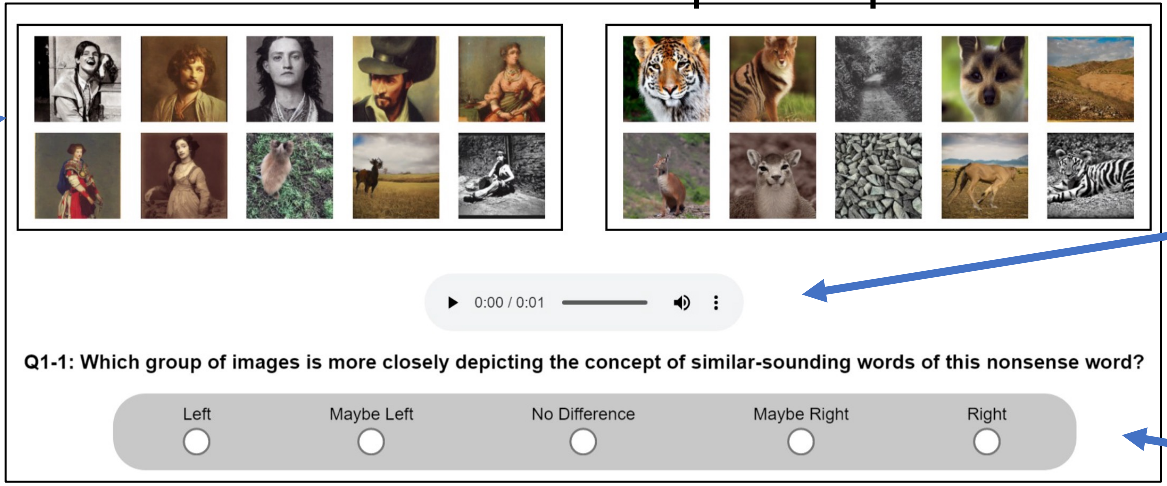
1. Reconstruct $L \times D_{\text{hidden}}$-dim. embedding from the D_{CLIP}-dim. one
   - Train a multilayer perceptron
2. Insert it into a pretrained Stable Diffusion model



Qualitative Evaluations

- Asked English speakers on **Amazon Mechanical Turk**
  - Two trials with different instructions
  - Trial 1: Choose *which images depict similar-sounding words?*
  - Trial 2: Choose *which images are more intuitive?*
  - Prepared 270 questions/nonwords from an English nonword dataset[7]

Interface shown to participants



10 generated images by either
- Proposed method
- Stable Diffusion (Comparative)

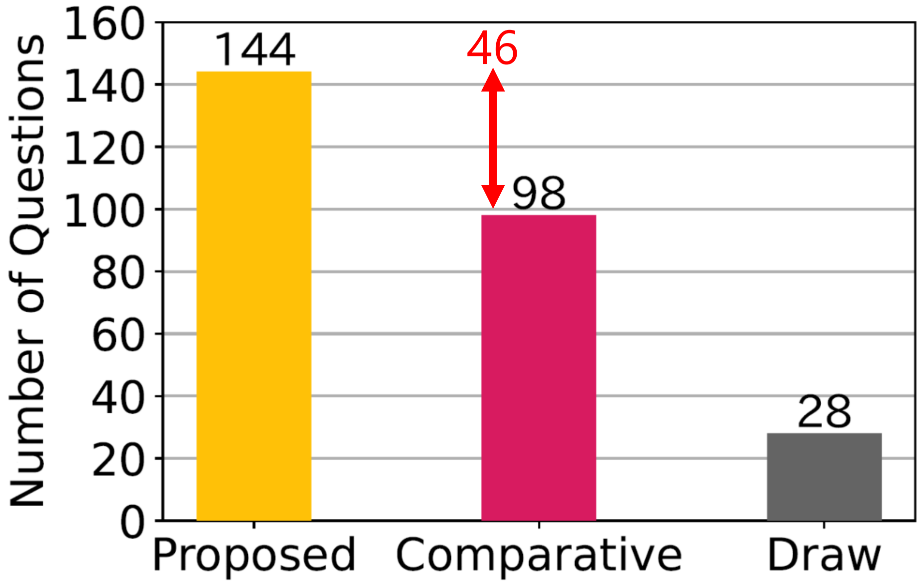
Audio pronouncing the nonword

Question

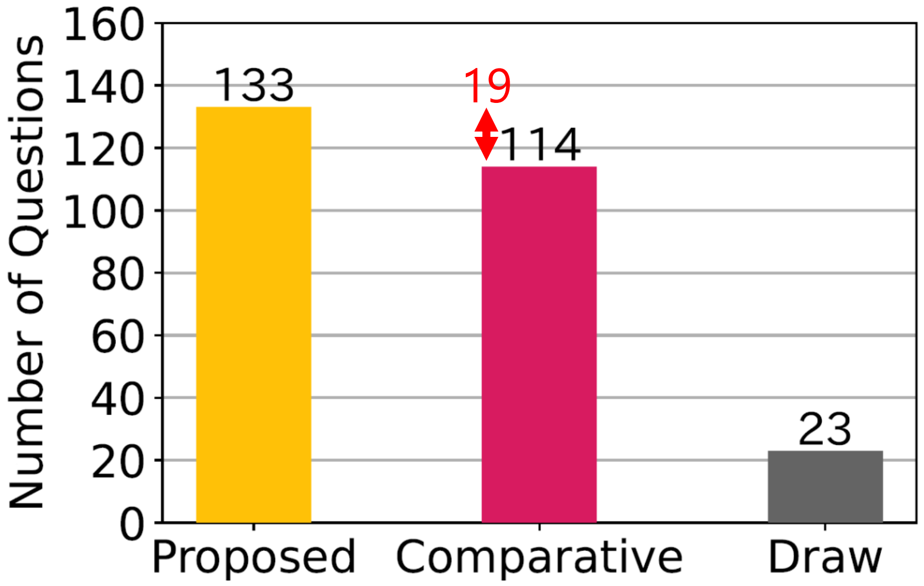
[7] Sabbatino et al., ""splink" is happy and "phrouth" is scary: Emotion intensity analysis for nonsense words.", WASSA 2022.

Results

- Proposed method wins over the comparative method
  - Generated images of the proposed method:
    - ✓Depict the concepts of their phonetically similar words more accurately
    - ✓Match human expectations more closely
- Proposed method has a larger gain in **Trial 1** than **Trial 2**
  - Intuitiveness involves more factors other than phonetic similarity



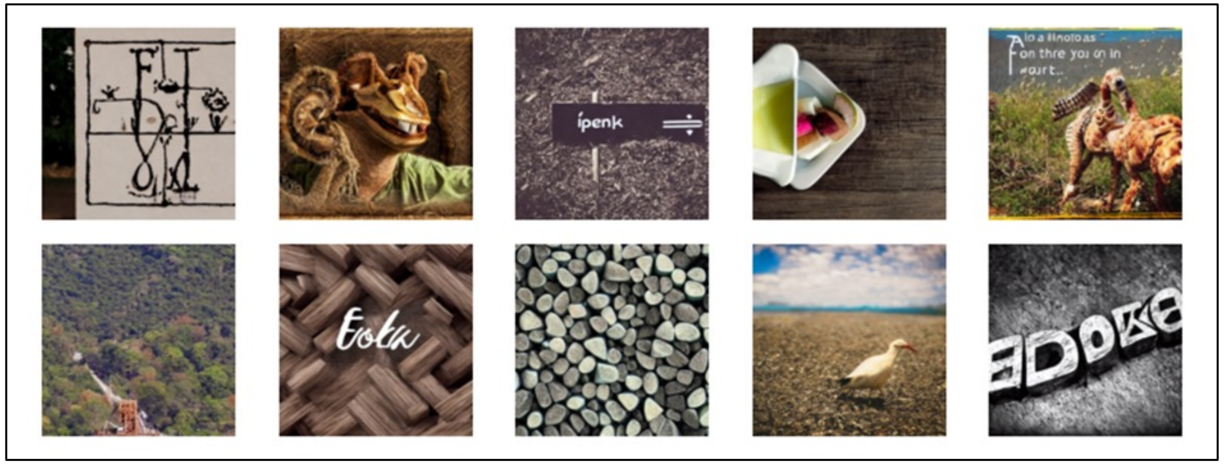
Trial 1: Contain similar-sounding words?



Trial 2: More intuitive?

Image Generation Example

What kind of imagery does **“Flike”** evoke in your mind?



Comparative (Stable Diffusion)

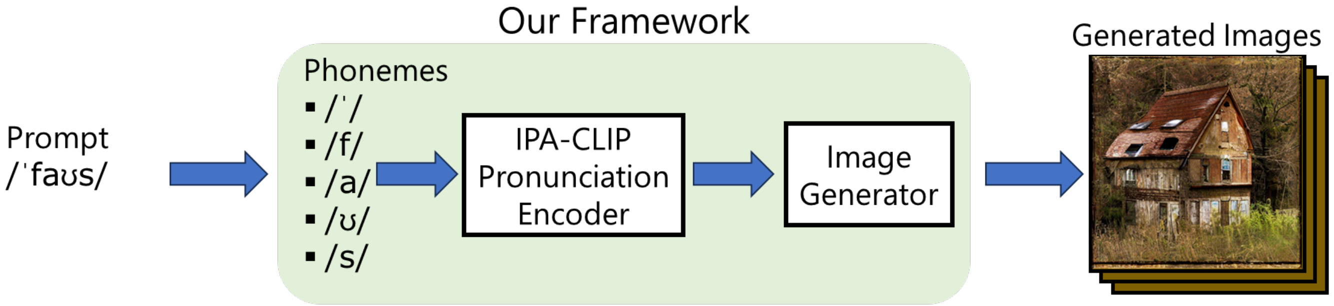
Proposed

Seemingly random
Not intuitive

Visual concept of flying or flight -> Bird
More intuitive!

Conclusion

- Pronunciation-to-Image generation robust against nonwords
  - Motivation: More intuitive nonword-to-image generation
  - Approach: Associate nonwords with their phonetically similar words
- Evaluation showed effectiveness of our method over Stable Diffusion
  - ✓ Depict phonetically similar (similar-sounding) words more accurately
  - ✓ Generate images more intuitive to humans



- Future Work
  - Extend to other languages and perform cross-lingual comparison
    - ➢ E.g., German, Japanese, and Chinese