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

Chihaya Matsuhira
matsuhirac@cs.is.i.nagoya-u.ac.jp
Nagoya University
Nagoya, Japan

Marc A. Kastner
mkastner@i.kyoto-u.ac.jp
Kyoto University
Kyoto, Japan

Takahiro Komamizu
taka-coma@acm.org
Nagoya University
Nagoya, Japan

Takatsugu Hirayama
t-hirayama@uhe.ac.jp
University of Human Environments
Okazaki, Japan

Keisuke Doman
kdoman@sist.chukyo-u.ac.jp
Chukyo University
Toyota, Japan

Ichiro Ide
ide@i.nagoya-u.ac.jp
Nagoya University
Nagoya, Japan

ABSTRACT

Text-to-Image (T2I) generation has long been a popular field of multimedia processing. Recent advances in large-scale vision and language pretraining have brought a number of models capable of very high-quality T2I generation. However, they are reported to generate unexpected images when users input words that have no definition within a language (nonwords), including coined words and pseudo-words. To make the behavior of T2I generation models against nonwords more intuitive, we propose a method that considers phonetic information of text inputs. The phonetic similarity is adopted so that the generated images from a nonword contain the concept of its phonetically similar words. This is based on the psycholinguistic finding that humans would also associate nonwords with their phonetically similar words when they perceive the sound. Our evaluations confirm a better agreement of the generated images of the proposed method with both phonetic relationships and human expectations than a conventional T2I generation model. The cross-lingual comparison of generated images for a nonword highlights the differences in language-specific nonword-imagery correspondences. These results provide insight into the usefulness of the proposed method in brand naming and language learning.

CCS CONCEPTS

• **Computing methodologies** → **Phonology / morphology**; • **Information systems** → *Multimedia and multimodal retrieval.*

KEYWORDS

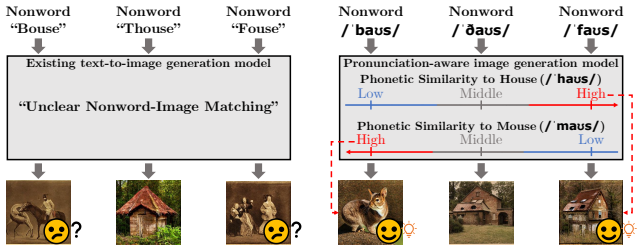
text-to-image generation, phonetics, psycholinguistics

ACM Reference Format:

Chihaya Matsuhira, Marc A. Kastner, Takahiro Komamizu, Takatsugu Hirayama, Keisuke Doman, and Ichiro Ide. 2023. Nonword-to-Image Generation Considering Perceptual Association of Phonetically Similar Words. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: New Methods and Practice (McGE '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3607541.3616818

McGE '23, October 29, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0278-5/23/10…$15.00
https://doi.org/10.1145/3607541.3616818



(a) Existing method [31] **(b) Proposed method**

Figure 1: Images generated by (a) an existing Text-to-Image (T2I) generation [31] and (b) the proposed pronunciation-aware image generation for three nonwords having various phonetic similarities to "*House*" and "*Mouse*". While the existing method mostly generates unrelated images for these nonwords, the proposed method draws either the concept of a house or a mouse depending on the phonetic similarity between the nonword and each of the two existing words.

1 INTRODUCTION

Text-to-Image (T2I) generation is the task of generating images that match a given text prompt. The recent innovation of large-scale multimodal pretraining has brought a number of T2I generation models performing very well [2, 19, 22, 25, 31], allowing us to easily generate plausible images that match a given text input. For instance, DALL·E [22] has shown a remarkable capability of T2I generation driven by a large-scale vision and language pretrained model called Contrastive Language-Image Pretraining (CLIP) [20]. Subsequently, a more lightweight and open-source latent diffusion model, Stable Diffusion [31], has been proposed which made it much easier for even non-experts to generate images for text inputs.

In psycholinguistics, it is known that even meaningless words (nonwords) can still evoke specific visual impressions in human minds [8]. For instance, hearing the sound of the pseudo-word "*Bouba*" tends to give the impression of a rounder shape than another

pseudo-word "*Kiki*" [12, 21]. These human-intrinsic nonword-to-imagery mappings are often used effectively in brand naming and language learning. For instance, names of characters and brands are designed so that the sounds of those names give impressions that match the visual characteristics of the target. Besides, since these sound-meaning correspondences are somewhat language-specific, being aware of their cross-lingual differences makes language learning much more effective. Existing T2I generation models, however, are not designed to capture these human intentions nor intuitions, causing cases where generated images do not match users' expectations [9] (See Fig. 1(a)).

The lack of criteria for processing nonwords can also increase the uncontrollability and vulnerability of language models [3, 18]. For example, it is reported that the English nonword "*Uccoisegeljaros*" can induce visual characteristics of birds in several English T2I generation models [3, 18]. This behavior not only seems unintuitive to humans but also could be used as a codeword to maliciously mislead these models to induce certain imagery.

To improve the robustness of the T2I generation models against nonwords and make their behavior more intuitive, we focus on the human nature of associating a nonword with its similar-sounding words. When English speakers hear a nonword "*Fouse*", they might recall its similar-sounding words such as "*House*" and "*Mouse*". Psycholinguistic findings suggest that the human brain recalls more phonetically similar words more quickly [5], indicating that the nonword "*Fouse*" would remind people of the word "*House*" more quickly and dominantly than "*Mouse*". Based on this, we construct a model by hypothesizing that (1) visual impressions evoked by any nonword are influenced by the visual characteristics of the concepts of phonetically similar words and (2) generating images containing the concept of a more phonetically similar word matches human expectations than those containing the concept of other words.

To this end, this paper proposes a method to generate images for a nonword considering phonetic similarity. To precisely describe the pronunciations of nonwords and calculate the phonetic similarity among words, the proposed method takes an array of International Phonetic Alphabet (IPA) symbols as an input. The pronunciations of existing words, which are used both in training and inference, are automatically converted using pronunciation dictionaries. The pronunciations of nonwords, which appear only in inference, will be specified by users, either by pronouncing the sound or more directly by typing phonetic symbols. Using IPA symbols, for instance, the existing word "*House*" is converted into its pronunciation /ˈhaʊs/. If a user expects the nonword "*Fouse*" to rhyme with "*House*", it will be transcribed as /ˈfaʊs/. The effect of phonetic similarity on the proposed image generation is illustrated in Fig. 1(b). Given a nonword /ˈfaʊs/, the proposed method generates images of its phonetically similar word "*House*" (/ˈhaʊs/) rather than "*Mouse*" (/ˈmaʊs/). Meanwhile, given another nonword /ˈbaʊs/, it generates images of "*Mouse*" rather than "*House*". This is because, according to phonetic features, the phoneme /f/ is more phonetically similar to /h/ than /m/ and the phoneme /b/ is more phonetically similar to /m/ than /h/. Such behavior requires knowledge of phonetic similarity and thus cannot be achieved just by replacing text inputs with pronunciation ones.

The proposed method is based on an existing T2I generation model, Stable Diffusion, which is a latent diffusion model [23] trained on a huge number of image-text pairs [26]. It employs the text encoder of CLIP to convert text inputs into conditioning vectors. We propose a method to substitute these text-based conditioning vectors with pronunciation-based ones obtained via a distillation approach [14], realizing pronunciation-aware image generation without retraining Stable Diffusion with additional training data.

2 RELATED WORK

2.1 Vision and Language Pretraining

The semantic gap between vision and language modalities has long been a primary concern in multimedia processing. Recently, using contrastive learning with a huge number of image-text pairs, OpenAI successfully trained Contrastive Language-Image Pretraining (CLIP) [20], which has achieved great success, especially in vision and language processing. CLIP consists of two encoders for two modalities: image and text encoders. The bimodal embedding space shared by the two encoders enables the similarity calculation between images and texts. This similarity calculation is powerful enough to even outperform state-of-the-art methods in several multimodal tasks without additional training.

So far, many CLIP extensions and applications have been proposed. Some researchers proposed encoders that map data of new modalities into the CLIP bimodal embedding space, such as multilingual texts [1], audio data [32], and phonetic transcription [14], enabling the similarity calculation among the image, the text, and the new modalities. The extension to phonetic transcription [14] considered the phonetic similarity between words with the aim of allowing users to input nonwords. Other studies applied CLIP to other vision-language tasks ranging from object detection [28] and image captioning [4] to text-to-image generation [2, 19, 22, 25, 31].

In the proposed method, the CLIP extension to phonetic transcription [14] is employed to convert a pronunciation input into a CLIP embedding vector. This enables nonwords to be input into the language model and the subsequent image generation model, thus realizing a nonword-to-image generation.

2.2 Recent Text-to-Image Generation Trend

The recent advance in Text-to-Image (T2I) generation owes greatly to the emergence of diffusion models [23, 29], with the help of large-scale pretrained models as introduced in Section 2.1. As opposed to conventional generative models like Generative Adversarial Networks (GANs) [6], which generate an image directly from a latent noise vector, diffusion models work by removing noise little by little from the noise vector until a clear image is obtained. Although this has resulted in many T2I generation models such as GLIDE [19], DALL·E 2 [22], and Imagen [25], these diffusion models had a problem of high computational costs required in their pixel-level denoising process. Latent diffusion models [23] solved this by applying the denoising procedure to latent vectors instead of images. This, combined with the CLIP text encoder, yielded an efficient but powerful T2I generation model called Stable Diffusion [31] which can be run even on a personal computer.

Yet, these models are not designed for text inputs containing nonwords, which could be problematic in some aspects. One of the concerns reported [9] is that, when users input nonwords, they

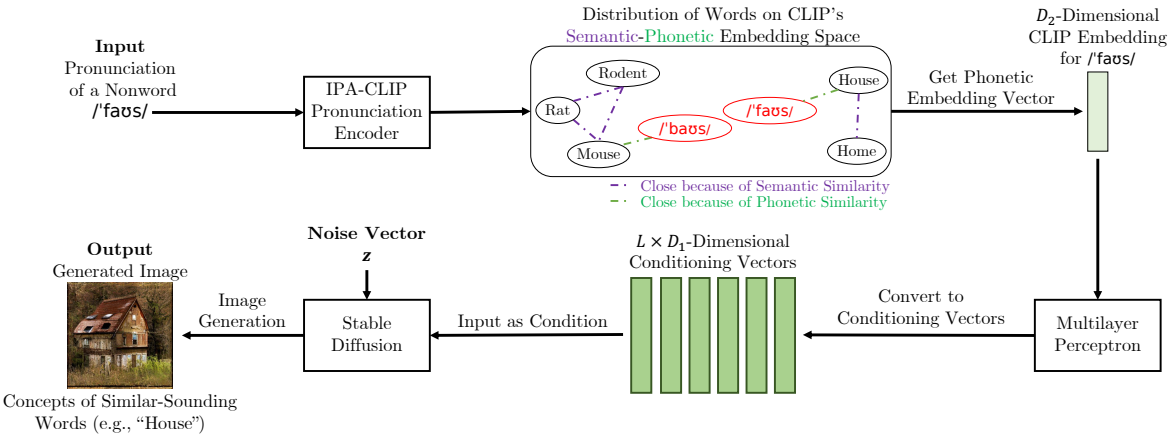


Figure 2: Framework of the proposed pronunciation-aware image generation using IPA-CLIP [14] and Stable Diffusion [31]. If a pronunciation of a nonword is input, it generates an image representing the concept of its phonetically similar word.

generally do not output images that match their expectations. Another problem is that the lack of criteria for processing nonwords could lead to an unintended use of these T2I generation models. Studies [3, 18] suggest that many models based on the CLIP text encoder implicitly associate certain nonwords with specific meanings. For example, Millière [18] proposed the nonword "*Uccoisegeljaros*" associated with the concept related to bird species. This is caused by the subword tokenization adopted in the text encoder [27] because it regards the nonword as a sequence of several subparts of words meaning birds in several major foreign languages (i.e., Italian "***Ucc**elli*", French "***Oise**aux*", German "*Vö**gel***", and Spanish "*Pá**jaros***"). Such a concatenation of subwords of foreign words meaning a specific concept can easily produce nonwords that can fool existing T2I generation models [18].

This paper controls Stable Diffusion by inserting conditioning pronunciation vectors computed based on phonetic similarity into a pretrained Stable Diffusion. This makes images generated for nonwords represent the concepts of their phonetically similar words, making the image generation more robust against nonwords.

3 IMAGE GENERATION FOR NONWORDS

The proposed method is built upon two existing methods: Stable Diffusion [31] and IPA-CLIP [14]. Stable Diffusion is an open-source latent diffusion model for Text-to-Image (T2I) generation [23], which was trained on a huge number of image-text pairs in Large-scale Artificial Intelligence Open Network (LAION) dataset [26]. For conditioning image generation on text prompts, it uses text features computed by the text encoder of Contrastive Language-Image Pretraining (CLIP) [20] which does not process nonwords in a phonetic manner. Hence, we replace this with IPA-CLIP, an extension of CLIP for a pronunciation input modality constructed by training an additional pronunciation encoder via distillation of the CLIP text encoder. Its pronunciation encoder maps a pronunciation written with International Phonetic Alphabet (IPA) into the same embedding space as CLIP. IPA-CLIP is selected because it calculates embedding vectors for input pronunciations, especially nonwords, based on pronunciation similarity defined in phonetics.

The framework of the proposed pronunciation-aware image generation is illustrated in Fig. 2. The proposed method assumes that the input text has been converted to IPA symbols in advance. Accordingly, a pronunciation and a noise vector are input to generate an image. To condition Stable Diffusion on the pronunciation input, we substitute the conditioning vectors of the CLIP text encoder used in Stable Diffusion with the pronunciation-based ones computed by IPA-CLIP. This approach requires much less computational costs and resources than retraining Stable Diffusion using pronunciation-image pairs.

3.1 Phonetic Embedding Vector Calculation

For the calculation of embedding vectors for pronunciation inputs, we use the framework of IPA-CLIP, which enables mapping from an array of phonetic symbols called International Phonetic Alphabet (IPA) to the text embedding space of CLIP. Its pronunciation encoder is trained via distillation of the CLIP text encoder using text-pronunciation pairs. For instance, given the input phoneme array /ə ˈfoʊˌtoʊ ˈʌv ə ˈkæt/ (IPA transcription for "a photo of a cat"), the pronunciation encoder learns to produce the embedding vector identical to the one calculated by the CLIP text encoder for the text prompt "`a photo of a cat`".

The phoneme tokenization method in IPA-CLIP considers articulatory phonetic features, which allows the model to calculate embedding vectors even for nonwords. Articulatory phonetic features of a phoneme refer to how the phoneme is pronounced. Consonants are assigned three features (place of articulation, manner of articulation, voicing), while vowels also have three other features (height of the tongue, backness of the tongue, roundedness of lips). Based on these features, for example, it is deduced that the consonant /b/ (bilabial, plosive, voiced) is more similar to /m/ (bilabial, **nasal**, voiced) than /h/ (**glottal**, **non-sibilant fricative**, **voiceless**). In contrast, because the features for vowels are considered continuous, the vowel /i/ (**high**, front, unrounded) is regarded as more similar to /e/ (**high-mid**, front, unrounded) than /a/ (**low**, front, unrounded), since the height feature ranges continuously between high and low.

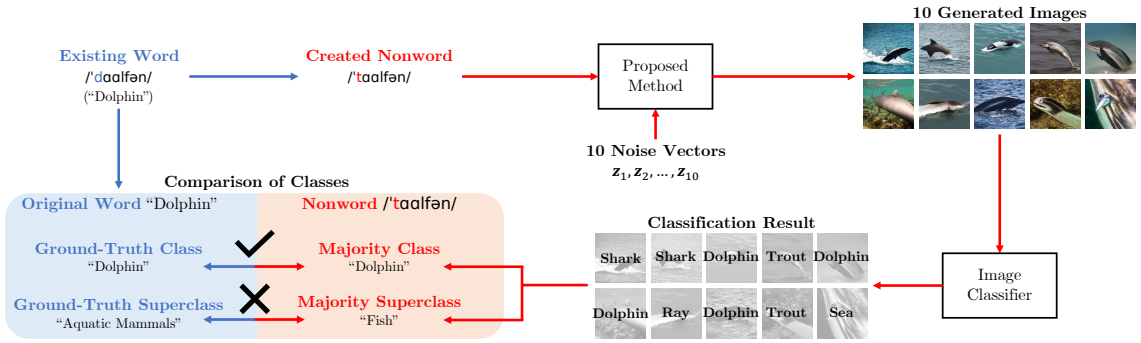


Figure 3: Evaluation procedure of the original-word-retrieval task using the existing word "*Dolphin*" and the nonword /ˈtɑalfən/ as an example.

By expanding this concept to the entire word, IPA-CLIP calculates an embedding vector for the pronunciation of a nonword so that the vector becomes close to those of its phonetically similar existing words. For example, given an input /ˈfaʊs/ (pronunciation of a nonword "*Fouse*"), its pronunciation encoder would compute a phonetic embedding vector similar to the one for "*House*" (/ˈhaʊs/) rather than "*Mouse*" (/ˈmaʊs/), since the pronunciation of "*Fouse*" is more phonetically similar to that of "*House*". At the same time, since words having a high semantic similarity to "*House*", such as "*Home*", are located close to "*House*" in the CLIP embedding space, the embedding vector of /ˈfaʊs/ may also be close to those of such semantically similar words. IPA-CLIP determines which phonetically similar words to approximate based on both pure phonetic similarity and word frequency, as this balance is learned via model training. For instance, even though the pronunciation /ˈfaʊs/ is phonetically similar to both "*House*" and "*Souse*" (/ˈsaʊs/), it would prefer "*House*" if it is a more frequently seen word in the training data. These characteristics of phonetic embedding vectors would also contribute to the image generation described in Section 3.2.

3.2 Inserting Phonetic Embedding Vector into Stable Diffusion

The text condition of Stable Diffusion requires an $L \times D_1$-dimensional last-hidden-state output of the Transformer in the CLIP text encoder (L is the maximum length of tokens of the Transformer), while the embedding vector calculated by IPA-CLIP is a D_2-dimensional final output of the CLIP text encoder. In the CLIP architecture, the D_2-dimensional output is computed from the $L \times D_1$-dimensional last-hidden-state output. Hence, to insert the embedding vectors of IPA-CLIP into Stable Diffusion, we need to perform this operation inversely and reconstruct $L \times D_1$-dimensional vectors from D_2-dimensional phonetic embedding vectors.

To learn this inverse function, this paper takes the straightforward strategy of training a multi-layer perceptron. It is trained using pairs of a D_2-dimensional final output and an $L \times D_1$-dimensional intermediate output of the original CLIP. After training this, we can obtain the last-hidden-state output vectors that correspond to the phonetic embedding vector of an input pronunciation. The obtained vectors are then inserted into Stable Diffusion along with a noise vector $\mathbf{z}$ to synthesize an image.

4 QUANTITATIVE EVALUATION

To quantitatively evaluate the influence of phonetically similar words on the proposed nonword-to-image generation, we perform an original-word-retrieval task of nonwords using the class names of CIFAR-100 image classification dataset [11]. Note that the term "*Nonword*" hereafter will denote the pronunciation of a non-existing word, and thus not the spelling (e.g., "a nonword /ˈfaʊs/").

To simplify the settings, this evaluation focuses only on the nonwords that surely have a certain phonetically similar word. Hence, given a nonword (e.g., /ˈtɑalfən/ in Fig. 3) created by slightly modifying the pronunciation of a certain existing word (e.g., "*Dolphin*" (/ˈdɑalfən/)), the goal is to find its original word from the contents of the generated images for the nonword. The core idea is that it is not always best to retrieve the similarly spelled original word. As described in Section 3.1, even if a nonword /ˈfaʊs/ originated from an existing word "*Mouse*" (/ˈmaʊs/), it should not generate images of a mouse for the nonword since there exists a more phonetically similar word "*House*" (/ˈhaʊs/). Thus, in this case, generating images of a house is regarded as more correct.

4.1 Task

The task of this evaluation is illustrated in Fig. 3. First, nonwords are prepared by slightly modifying the class names of the CIFAR-100 dataset. This dataset provides 100 classes of visually distinguishable common objects ranging from animals to furniture along with twenty superclasses consisting of five classes each. Hence, each nonword has both a ground-truth class and a superclass derived from the original word. For instance, the nonword /ˈtɑalfən/ is assigned with the ground-truth class "*Dolphin*" and also its superclass "*Aquatic Mammals*". Note that the two superclasses identically named "*Vehicles*" are merged into one in this evaluation, resulting in nineteen unbalanced superclasses in total.

Next, the proposed method generates ten images for each nonword. On these generated images, we then perform 100-class image classification based on CIFAR-100 classes to estimate the object that most frequently appears in them. For example, if the majority of the generated images for /ˈtɑalfən/ are predicted to contain dolphins and are classified as "*Dolphin*", the majority class for /ˈtɑalfən/ is regarded as the class "*Dolphin*". At the same time, we also calculate the majority superclass for /ˈtɑalfən/ as the most frequent

superclass among the ten predicted superclasses in order to assess whether the generated images for the nonword depict at least dolphin-like concepts, even if they are not exactly dolphins.

Finally, the majority class/superclass and that of the original word are compared to calculate accuracy. Observing the transition of accuracy according to the phonetic similarity between a nonword and its original word measures how well the proposed method associates a nonword with its phonetically similar word.

4.2 Nonword Creation

Nonwords are created in two schemes: Consonant substitution and vowel substitution. In the former scheme, nonwords are created by replacing the initial consonant of each of the CIFAR-100 classes with other English consonants. For example, from the class name "*Dolphin*" (/ˈdɑɑlfən/), we obtain nonwords such as /ˈtɑɑlfən/ ("*Tolphin*") and /ˈgɑɑlfən/ ("*Golphin*"). In this scheme, we only use classes that start from a single consonant, yielding 936 nonwords stemming from 68 existing words. In the latter scheme, nonwords are created by replacing the vowel of the first syllable in each of the CIFAR-100 classes with other English monophthongs or diphthongs. For example, from the class name "*Dolphin*" (/ˈdɑɑlfən/), we obtain nonwords such as /ˈdoʊlfən/ ("*Dhole*"+"*-phin*") and /ˈdɔɪlfən/ ("*Doyle*"+"*-phin*"). Considering English phonological restrictions, we replace checked vowels only with checked vowels while we replace free vowels only with free vowels. In this scheme, we use classes that start from one or more consonants, yielding 369 nonwords stemming from 77 existing words.

Next, we define the phonetic similarity between each original word-nonword pair based on the articulatory phoneme features as described in Section 3.1. For those in the consonant substitution scheme, only the initial consonants are different. Hence, following the previous work [14], the phonetic similarity is defined using the number of common features out of the three consonant articulatory features in the two contrasting consonants. The similarity is regarded as "High" if the number is two, "Middle" if it is one, and "Low" if no feature is in common. For those in the vowel substitution scheme, we first construct a 3-dimensional vowel space using the three vowel articulatory features, as shown in Fig. 4. We then define the phonetic similarity by computing the L1 distance d between the two vowels on the space. Here, if the vowel is a diphthong, the geometric centroid is used as the point of the diphthong. For instance, when comparing /ɑɑ/ and /oʊ/, d will be the L1 distance between the points $(1, 1, 1)$ and $(1, \frac{1}{6}, 1)$. The similarity is regarded as "High" if $d \leq 1$, "Middle" if $1 < d \leq 2$, and "Low" if $2 < d$.

To compare the proposed method with the text-based original Stable Diffusion [31], we also prepare corresponding spellings for each nonword. In this paper, the phoneme-spelling correspondence is decided based on the co-occurrences of phonemes and spellings in existing English words (See Appendix A.1 for details). According to this, some nonwords will be spelled exactly the same as the original word (e.g., "*Dolphin*" for the nonword /ˈdoʊlfən/), which may yield situations advantageous to the comparative method in identifying the original word. Finally, the Carnegie Mellon University (CMU) dictionary[1] is used to check if the pronunciations of created nonwords do not exist in English. Note that this may not cover

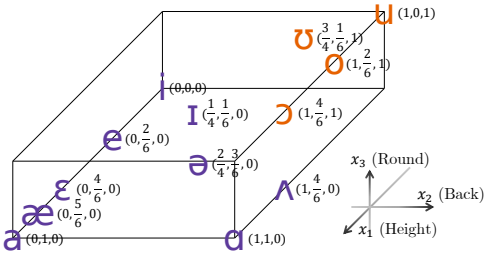


Figure 4: Vowel space used to measure the phonetic similarity among vowels.

Table 1: Statistics of nonwords in each experimental setting.

Consonant Substitution	Low	Middle	High
Number of nonword	324	424	188
Example nonword for /ˈdɑɑlfən/ – Spelling (Nonword for "*Dolphin*")	/ˈhɑɑlfən/ "*Holphin*"	/ˈkɑɑlfən/ "*Colphin*"	/ˈtɑɑlfən/ "*Tolphin*"
Vowel Substitution	**Low**	**Middle**	**High**
Number of nonwords	81	191	97
Example nonword for /ˈtuuɪps/ – Spelling (Nonword for "*Tulips*")	/ˈtaɪɪps/ "*Tilips*"	/ˈtɔɪɪps/ "*Toilips*"	/ˈtoʊɪps/ "*Tolips*"

all inflected forms of words and rare words, and thus the created nonwords might contain the pronunciations of such words. Table 1 shows the statistics of the created nonwords in each scheme.

4.3 Experimental Settings

4.3.1 Proposed and Comparative Methods. The proposed method requires IPA-CLIP [14] distilled from the CLIP ViT-L/14 model [20]. To adjust the model to our purpose including the image generation and the multilingual comparison, we prepare pronunciation-text pairs by ourselves and retrain its pronunciation encoder from scratch. Specifically, first, we change the pronunciation dictionary from the original paper [14], which is used to convert words into phonetic transcription, to the CMU dictionary[1]. Also, the phonemes /ʧ/ and /ʤ/ are split into combination of two separate ones, /tʃ/ and /dʒ/, respectively. For text data, we use 1,114,375 English sentences taken from some image captioning datasets [1], as well as 26,143 sentences consisting of only one word from Spell Checker Oriented Word Lists (SCOWL)[2]. In addition, to emphasize the existence of an indefinite article before words, we also include sentences in the shape of "`a photo of <WORD>`" and "`a photo of a <WORD>`". Sentences that have less frequent words are removed using a Python package `wordfreq` [30]. IPA-CLIP is trained up to 100 epochs with these 1,192,804 text-pronunciation pairs while applying the same settings for the other hyperparameters as the original paper. For Stable Diffusion, we use `Stable Diffusion-v1-4` with the classifier-free guidance scale of 7.5. Ten images are generated for each nonword using the same noise vectors.

Two-layer perceptron with the ReLU activation is adopted as the multi-layer perceptron of the proposed method. It is trained up to 1,000 epochs using CLIP text embedding vectors and corresponding

[1]https://github.com/menelik3/cmudict-ipa/ (Accessed July 11, 2023)

[2]http://wordlist.aspell.net/ (Accessed July 11, 2023)

last-hidden-state outputs of sentences included in the training data of IPA-CLIP. Mean Squared Error is used as an objective function.

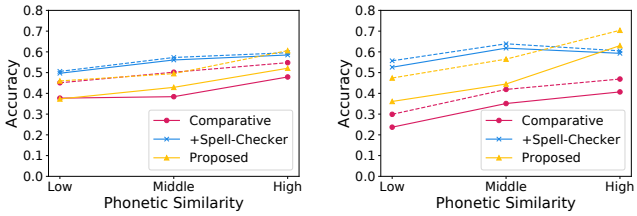
As a comparison method, we first adopt the raw Stable Diffusion. It generates images from the spelling of a nonword. We do not further retrain its weights. We also compare another simple strategy for nonword inputs which we call a Spell-Checker strategy. This strategy replaces the nonword in a text prompt with the most frequent existing word among words having the smallest edit distance to the nonword before inputting it to Stable Diffusion. For example, given a nonword "*Bouse*", it first finds a set of words that have the smallest edit distance to the nonword, e.g., {"*Mouse*", "*House*", "*Blouse*"}. It then selects the most frequent word among them, "*House*", regardless of the phonetic similarity, and inputs the prompt "`a photo of a house`" to Stable Diffusion to obtain images for the nonword "*Bouse*". `wordfreq` is used for the frequency calculation, and SCOWL is used as a source of existing words.

4.3.2 Prompt Engineering in Image Generation. Prompt engineering is a modification added to the language input of a pretrained model to specify the context and intention of the input. Following the prompt engineering for CLIP, the prompt for the proposed method is set as /ə ˈfoʊˌtoʊ ˈʌv ə <NONWORD>/ (also "`a photo of a <NONWORD>`" for the comparative text-based Stable Diffusion) to restrict the domain of generated images. For instance, in the example shown in Fig. 3, ten images are generated from the prompt /ə ˈfoʊˌtoʊ ˈʌv ə ˈtɑɑlfən/ instead of just /ˈtɑɑlfən/. If a one-word prompt /ˈtɑɑlfən/ is used, the ten generated images would become too diverse and make it hard to obtain the majority class since the word /ˈtɑɑlfən/ itself has no explicit meaning.

4.3.3 CIFAR-100 Image Classifier. CLIP is employed as an image classifier of CIFAR-100 classes, as used in the original paper [20]. CLIP image classifier functions based on the similarity calculation between a given image and each of the class names. Given an image, CLIP first computes its image embedding and also the text embeddings for all of the class names. Then, the cosine similarity between the image embedding and each of the text embeddings is calculated. Softmaxed similarity scores work as a class probability distribution. Hence, CLIP chooses the class that gives the maximum similarity to the image embedding as the predicted class. As prompt engineering, we use "`a photo of a <CLASS>`" if the class is a singular noun, while "`a photo of <CLASS>`" if it is plural.

4.4 Results and Discussions

The results in the consonant and vowel substitution schemes are shown in Fig. 5. In both schemes, the Spell-Checker approach made the performance of the comparative method much higher, resulting in an always high accuracy regardless of the phonetic similarity between original words and nonwords. In contrast, the proposed method shows a clear tendency that accuracy correlates with the phonetic similarity between original words and nonwords, and outperforms the others when the phonetic similarity is high. This means that the more phonetically similar a nonword is to its original word, the more likely generated images are to contain the concept of its original words, indicating that the proposed method captures the phonetic relationships among phonemes more correctly than the comparative methods. Interestingly, the raw comparative method also showed this tendency, although it was less clear.



(a) Accuracy for 936 nonwords in the consonant substitution scheme.

(b) Accuracy for 369 nonwords in the vowel substitution scheme.

Figure 5: Results of the original-word-retrieval task. In each figure, the solid lines represent the accuracy for classes, while the dashed lines represent the accuracy for superclasses of CIFAR-100 [11].

Regarding the proposed method, we also confirmed that the accuracy for superclasses was always higher than that for classes by more than $\frac{20}{19\times19} \approx 0.06$ which is the chance rate for the superclass prediction in this evaluation. This implies, even if the generated images from a nonword (e.g., /ˈkɑɑlfən/) do not contain the very concept of its original word ("*Dolphin*"), they tend to at least depict similar concepts of the original word (e.g., "*Whale*"), which could simulate the activation of a human brain towards semantically similar words [15]. This tendency was not observed when using the Spell-Checker approach because it actually generates images from text prompts that do not contain nonwords.

In the consonant substitution, although the proposed method overall outperformed the raw comparative method, a great improvement was not confirmed. Language models such as Word2vec [16, 17] are known to learn phonetic relationships among phonemes from the phonological restrictions on phoneme cooccurrences in English [10]. We assume that the CLIP text encoder could also have learned such relationships, resulting in the relatively high performance of the raw Stable Diffusion in this scheme. In the vowel substitution, on the other hand, the proposed method always performed significantly better than the comparative method. One reason is the lack of means to precisely describe vowels in the English language. As different vowels may sometimes be spelled with the same letters unlike consonants, the CLIP text encoder as well as Stable Diffusion could not learn the vowel phonetic features implicitly from the huge number of texts.

To further discuss the behavior of the proposed method towards consonants, we conducted an additional evaluation that measures its performance for unseen foreign consonants. As some foreign consonants share certain articulatory features with English ones, the proposed method can generate images even from inputs containing such consonants. For instance, the German phoneme /ç/ (as in "***Ich***") has the same manner of articulation as /f/ (Fricative), hence the proposed method may be able to associate the nonword /ˈçaɪt/ with, e.g., "*Fight*" (/ˈfaɪt/). The setting of this evaluation is the same as that of the consonant substitution, except for the nonwords used. We prepare nonwords by substituting the initial

Table 2: Statistics of nonwords in the additional evaluation.

Foreign Consonant Substitution	Low	Middle	High
Number of nonwords	500	670	204
Example nonword for /ˈdɑɑlfən/	/ˈçɑɑlfən/	/ˈɲɑɑlfən/	/ˈɟɑɑlfən/
– Description of the initial phone	German 'ch'	Spanish 'ñ'	Czech 'ď'

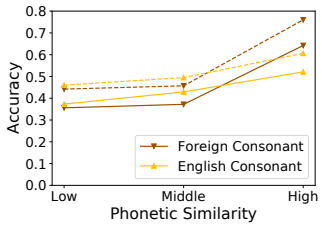


Figure 6: Accuracy of the original-word-retrieval task for 1,374 nonwords in the foreign-consonant substitution scheme using the proposed method. For comparison, we also plot its accuracy in the English consonant substitution scheme shown in Fig. 5(a).

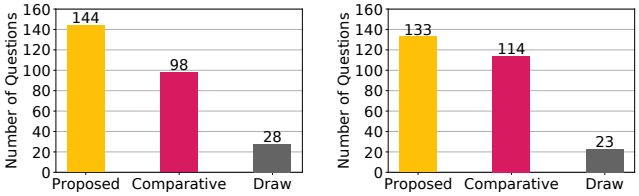
consonant of each class with either of twenty foreign consonants. This scheme resulted in 1,374 nonwords stemming from 71 existing words. The statistics are shown in Table 2.

The result of this additional evaluation is shown in Fig. 6. When compared with Fig. 5(a), accuracy has increased particularly for nonwords whose phonetic similarity to their original words is "High". As a result, the proposed method scored an accuracy of 0.760 for superclasses when nonwords have a high phonetic similarity to their original words, meaning that it can associate 76.0% of nonwords with the concepts of their phonetically very similar words if such words exist. One of the reasons for this increase is that nonwords starting from a non-English consonant have fewer phonetically very similar words than nonwords starting from an English consonant, making them more likely to be associated with their original words. For example, the nonword /ˈtɑɑlfən/ may be phonetically similar to both "*Dolphin*" and "*Tall*", while the nonword /ˈɟɑɑlfən/ may not be considered phonetically similar to "*Tall*".

In summary, we have confirmed that the proposed method associates nonwords with the concepts of their phonetically similar words based on phonetic similarity, even for nonwords containing unseen phonemes. This is the result of introducing the pronunciation modality and thus cannot be achieved using conventional text-based approaches, especially the Spell-Checker approach which totally ignores the phoneme relationships.

4.5 Limitations

Although the proposed method outperformed Stable Diffusion regarding nonwords, the drop in its general performance towards existing words was also confirmed. We observed an increase of Fréchet Inception Distance (FID) [7] by 6.4 points and a decrease of CLIP score [20] by 0.023 points on Microsoft Common Objects in Context (MS-COCO) validation dataset [13]. We believe this is caused by both the small model size of IPA-CLIP compared to CLIP and the simple embedding insertion strategy to Stable Diffusion.



(a) Result of the first trial. **(b) Result of the second trial.**

Figure 7: Number of questions in which the images generated by each method were preferred in the two trials of the qualitative evaluation. The first trial asked participants "which method generates images of similar-sounding words of a given nonword", while the second trial asked "which method generates images that intuitively match the nonword".

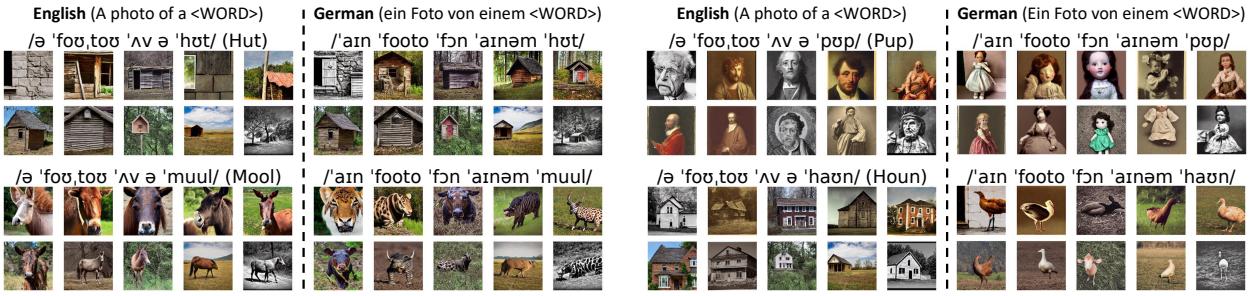
For this, combining both text and pronunciation modalities would also enhance image generation performance.

5 QUALITATIVE EVALUATION

To further evaluate the proposed nonword-to-image generation for other types of nonwords, we conduct a user study on Amazon Mechanical Turk[3]. In each question of the survey, given an audio file pronouncing a nonword and two groups of ten generated images, one for the proposed method and the other for the comparative method (Stable Diffusion [31]), five participants are asked to answer which group of images matches the sound of the nonword. We conduct two trials with different instructions. In the first trial, 73 participants are instructed to choose "the group that more closely depicts the concept of words similar-sounding to the nonword". In the second trial, 108 participants are instructed to choose "intuitively the group that more closely depicts the sound of the nonword". In both trials, we use the same set of 270 randomly created English nonwords taken from a dataset of nonwords annotated with evoked emotion labels [24]. Hence, in contrast to the quantitative evaluation, these nonwords do not necessarily have specific phonetically similar words. See Appendix A.2 for more detailed experimental settings.

The results of the two trials are shown in Fig. 7. In both trials, the proposed method was preferred in more questions than the comparative method. These results indicate that, even for randomly created nonwords, the images generated by the proposed method depict the concepts of their phonetically similar words more accurately, and they also match human expectations more than those generated by the comparative method. The latter result reinforces our hypothesis that generating the concept of a phonetically similar word for a nonword matches human expectations more than generating the concepts of other words. Nevertheless, in the second trial, the gain of the proposed method was not as large as in the first trial. One of the reasons is that the association of phonetically similar words is not the sole factor that shapes human perception. To generate more intuitive images for nonwords, collecting more human annotations and applying other psycholinguistic findings could be effective.

[3]https://www.mturk.com/ (Accessed July 11, 2023)



(a) Examples of nonwords from which images containing similar concepts are generated between English and German.

(b) Examples of nonwords from which images containing different concepts are generated between English and German.

Figure 8: Cross-lingual comparison of the proposed nonword-to-image generation models for English and German.

6 CROSS-LINGUAL COMPARISION

In the previous sections, the proposed method was constructed solely for English. This section explores its applicability to other languages and its usefulness in comparing visual concepts evoked by nonwords among different languages. As such, we train the proposed method for German, which is selected because it has a similar phoneme/phone set to the English one, and seek which pronunciations evoke similar/different imagery in the two languages. For instance, the German word "*Gift*" (Poison) has an opposite meaning but has the same pronunciation as the English word "*Gift*", and thus the nonwords phonetically similar to the pronunciation of "*Gift*" may evoke different visual concepts between the speakers of these languages. If the proposed method can visualize such differences, it would be useful in brand naming and language learning.

We first prepare German sentence-pronunciation pairs and train IPA-CLIP [14] on the German data, which is used to condition Stable Diffusion [31] with German pronunciation inputs in the proposed method (See Appendix A.3 for details). Next, as seeds of the nonword-to-image generation, 511 monosyllabic nonwords in the form of Consonant-Vowel-Consonant are prepared. These nonwords do not exist in both English and German and consist of only phonemes appearing in both languages. We generate images using a prompt /ə ˈfoʊˌtoʊ ˈʌv ə <NONWORD>/ ("A photo of a <NONWORD>") for English and /ˈaɪn ˈfooto ˈfɔn ˈaɪnəm <NONWORD>/ ("Ein Foto von einem <NONWORD>") for German.

Figure 8 shows examples of nonwords from which images containing similar or dissimilar concepts were generated between English and German. For the nonword /ˈhʊt/, for example, concepts related to a hut appeared in both languages. This is because the pronunciation of the nonword is phonetically similar to both English "*Hut*" (/ˈhʌt/) and German "*Hütte*" (/ˈhʏtə/, a German word for hut). In contrast, between the two languages, different genders of people appeared in the generated images for the nonword /ˈpʊp/. In this case, the English "*Pope*" (/ˈpoʊp/) and the German "*Puppe*" (/ˈpʊpə/, a word for doll) has appeared in those generated images, as they are located quite close to the nonword /ˈpʊp/ in the embedding space of IPA-CLIP in each language.

From the results above, we confirmed that the characteristic of associating nonwords with their phonetically similar words is preserved in the model built for the German language, too. This assures the applicability of the proposed method to at least languages having a phonological rule similar to English. Modeling such language-specific nonword-to-imagery mappings should be useful in applications like brand naming and language learning.

7 CONCLUSIONS

Conventional Text-to-Image (T2I) generation models are reported to generate unexpected images when nonwords are input [3, 9, 18]. To make their behavior more robust against nonwords, this paper proposed a method to generate images considering phonetic similarity. This enables the model to connect nonwords with their phonetically similar existing words. The quantitative evaluation showed its ability to generate images that contain concepts of phonetically similar words for nonwords better than a conventional T2I generation method. This was even valid when nonwords contained unseen and foreign phonemes. The qualitative evaluation further confirmed a better agreement of the proposed nonword-to-image generation with human expectations using a wide variety of nonwords. Lastly, the comparison of the proposed models built for different languages suggested its usefulness in brand naming and language learning.

Future work includes further analysis of the nonword-image matching in the proposed method. Also, we recognize that the proposed framework does not distinguish homonyms and generates the same images for such words. For this, combining phonetic and morphological similarities would enable applying the proposed method also to languages rich in homonyms such as Japanese.

ACKNOWLEDGMENTS

This work was partly supported by Microsoft Research CORE16 program and JSPS Grant-in-aid for Scientific Research (22H03612). This work is a result of a joint research project between Nagoya University and the National Institute of Informatics. The computation was carried out using the General Projects on supercomputer "Flow" at Information Technology Center, Nagoya University.

The first author would like to take this opportunity to thank the "Nagoya University Interdisciplinary Frontier Fellowship" supported by Nagoya University and JST, the Establishment of University Fellowships towards the Creation of Science Technology Innovation, Grant Number JPMJFS2120.

REFERENCES

[1] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proc. 13th Lang. Resour. Evaluation Conf.* (Marseille, Bouches-du-Rhône, France). ELRA, Paris, France, 6848–6854.

[2] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. VQGAN-CLIP: Open domain image generation and editing with natural language guidance. Comput. Res. Reposit., arXiv Preprint, arXiv:2204.08583. https://doi.org/10.48550/arXiv.2204.08583

[3] Giannis Daras and Alexandros G. Dimakis. 2022. Discovering the hidden vocabulary of DALLE-2. In *Proc. NeurIPS 2022 Workshop Score-Based Methods* (New Orleans, LA, USA). 5 pages. https://openreview.net/forum?id=jxeSZaVzpmg

[4] Federico Galatolo, Mario Cimino, and Gigliola Vaglini. 2021. Generating images from caption and vice versa via CLIP-guided generative latent space search. In *Proc. Int. Conf. Image Process. Vis. Eng.* (Prague, Czech). SciTePress, Setúbal, Portugal, 166–174. https://doi.org/10.5220/0010503701660174

[5] Stephen Goldinger, Paul Luce, David Pisoni, and Joanne Marcario. 1992. Form-based priming in spoken word recognition: The roles of competition and bias. *J. Exp. Psychol. Learn. Mem. Cogn.* 18, 6 (1992), 1211–1238. https://doi.org/10.1037/0278-7393.18.6.1211

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Adv. Neural Inf. Process. Syst.* (Montréal, QC, Canada), Vol. 27. Curran Associates, Inc., New York, NY, USA, 9 pages.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Adv. Neural Inf. Process. Syst.* (Long Beach, CA, USA), Vol. 30. Curran Associates, Inc., New York, NY, USA, 12 pages.

[8] Leanne Hinton, Johanna Nichols, and John J. Ohala. 1995. *Sound Symbolism.* Cambridge University Press, Cambridge, England, UK. https://doi.org/10.1017/CBO9780511751806

[9] Hyung-Kwon Ko, Gwanmo Park, Hyeon Jeon, Jaemin Jo, Juho Kim, and Jinwook Seo. 2023. Large-scale text-to-image generation models for visual artists' creative works. In *Proc. 28th Int. Conf. Intell. User Interfaces* (Sydney, NSW, Australia). ACM, New York, NY, US, 919–933.

[10] Sudheer Kolachina and Lilla Magyar. 2019. What do phone embeddings learn about Phonology?. In *Proc. 16th Workshop Comput. Res. Phonetics, Phonol., Morphol.* (Firenze, Toscana, Italy). ACL, Stroudsburg, PA, USA, 160–169. https://doi.org/10.18653/v1/W19-4219

[11] Alex Krizhevsky. 2009. *Learning Multiple Layers of Features from Tiny Images.* Master's thesis. University of Toronto, Toronto, ON, Canada. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf

[12] Wolfgang Köhler. 1929. *Gestalt Psychology.* H. Liveright, New York, NY, USA.

[13] Tsung Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Ramanan Deva, Dollár Piotr, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proc. 13th Europ. Conf. Comput. Vis. Part V* (Zurich, Switzerland). Springer, Cham, Basel, Switzerland, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

[14] Chihaya Matsuhira, Marc A. Kastner, Takahiro Komamizu, Takatsugu Hirayama, Keisuke Doman, Yasutomo Kawanishi, and Ichiro Ide. 2023. IPA-CLIP: Integrating phonetic priors into vision and language pretraining. Comput. Res. Reposit., arXiv Preprint, arXiv:2303.03144. https://doi.org/10.48550/arxiv.2303.03144

[15] David Meyer and Roger Schvaneveldt. 1971. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *J. Exp. Psychol.* 90, 2 (11 1971), 227–234. https://doi.org/10.1037/h0031564

[16] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. Comput. Res. Reposit., arXiv Preprint, arXiv:1301.3781. https://doi.org/10.48550/arXiv.1301.3781

[17] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Adv. Neural Inf. Process. Syst.* (Lake Tahoe, NV, USA), Vol. 26. Curran Associates, Inc., New York, NY, USA, 3111–3119.

[18] Raphaël Millière. 2022. Adversarial attacks on image generation with made-up words. Comput. Res. Reposit., arXiv Preprint, arXiv:2208.04135. https://doi.org/10.48550/arXiv.2208.04135

[19] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proc. 39th Int. Conf. Mach. Learn., Proc. Mach. Learn. Res.* (Baltimore, MD, USA), Vol. 162. PMLR, Cambridge, MA, USA, 16784–16804.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proc. 38th Int. Conf. Mach. Learn., Proc. Mach. Learn. Res.* (Online), Vol. 139. PMLR, Cambridge, MA, USA, 8748–8763.

[21] Vilayanur S. Ramachandran and Edward M. Hubbard. 2001. Synaesthesia —A window into perception, thought and language. *J. Conscious. Stud.* 8, 12 (2001), 3–34.

[22] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with CLIP latents. Comput. Res. Reposit., arXiv Preprint, arXiv:2204.06125. https://doi.org/10.48550/arXiv.2204.06125

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proc. 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (New Orleans, LA, USA). IEEE, New York, NY, USA, 10684–10695.

[24] Valentino Sabbatino, Enrica Troiano, Antje Schweitzer, and Roman Klinger. 2022. "splink" is happy and "phrouth" is scary: Emotion intensity analysis for nonsense words. In *Proc. 12th Workshop Comput. Approaches to Subj. Sentiment Soc. Media Anal.* (Dublin, Ireland). ACL, Stroudsburg, PA, USA, 37–50. https://doi.org/10.18653/v1/2022.wassa-1.4

[25] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. Comput. Res. Reposit., arXiv Preprint, arXiv:2205.11487. https://doi.org/10.48550/arxiv.2205.11487

[26] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W. Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R. Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Proc. 36th Conf. Neural Inf. Process. Syst. Datasets Benchmarks Track* (New Orleans, LA, USA). Curran Associates, Inc., New York, NY, USA, 17 pages.

[27] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. 54th Annual Meet. Assoc. Comput. Linguist.* (Berlin, Germany), Vol. 1. ACL, Stroudsburg, PA, USA, 1715–1725. https://doi.org/10.18653/v1/P16-1162

[28] Hengcan Shi, Munawar Hayat, Yicheng Wu, and Jianfei Cai. 2022. ProposalCLIP: Unsupervised open-category object proposal generation via exploiting CLIP cues. In *Proc. 2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* (New Orleans, LA, USA). IEEE, New York, NY, USA, 9611–9620.

[29] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. 32nd Int. Conf. Mach. Learn., Proc. Mach. Learn. Res.* (Lille, Nord, France), Vol. 37. PMLR, Cambridge, MA, USA, 2256–2265.

[30] Robyn Speer, Joshua Chin, Andrew Lin, Sara Jewett, and Lance Nathan. 2018. LuminosoInsight/wordfreq: v2.2. Zenodo. https://doi.org/10.5281/zenodo.1443582

[31] Computer Vision and Learning Research Group at Ludwig Maximilian University of Munich. 2022. Stable Diffusion. https://github.com/CompVis/stable-diffusion/ (Accessed July 11, 2023).

[32] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2CLIP: Learning robust audio representations from CLIP. In *Proc. 2022 IEEE Int. Conf. Acoust. Speech Signal Process.* (Singapore). IEEE, New York, NY, USA, 4563–4567. https://doi.org/10.1109/ICASSP43922.2022.9747669

A APPENDIX

A.1 Phoneme-Spelling Correspondence

For each of the nonwords created in Section 4.2 (e.g., /ˈtɑɑlfən/), the paper also prepared a corresponding spelling (e.g., "*Tolphin*"). Table 3 and Table 4 show the phoneme-spelling correspondences for consonants and vowels, respectively. These correspondences were defined based on the spellings and pronunciations of existing English words. For instance, 'u' was selected as the spelling for the phoneme /ʌ/ because, in most of the words having the phoneme /ʌ/, its corresponding spelling is written using 'u', such as "*C**u**t*" (/ˈkʌt/), "*S**u**m*" (/ˈsʌm/), and "***U**tter*" (/ˈʌtəɹ/).

A.2 Qualitative Evaluation

This section elaborates on the experimental settings of the qualitative evaluation described in Section 5.

A.2.1 Crowdsourcing. We recruited 73 English speakers living in the United States on Amazon Mechanical Turk (AMT)[3]. Given an audio file pronouncing a nonword and two groups of ten generated images; one for the proposed method and the other for the comparative method (original Stable Diffusion [31]), participants were asked to answer which group of images matches the sound of the nonword. Specifically, they were instructed to repeat the nonword to familiarize themselves with the pronunciation and then choose "the group that more closely depicts the concept of words that sound similar to the nonsense word". For each task (called "*HIT*" in AMT), eleven sets of questions were sequentially shown to participants, including ten actual questions and one attention check question which should be answered correctly. For each question, participants were also asked to write what kinds of similar-sounding words they recalled. These data were used only for rejecting inappropriate answers. The example of the user interface is shown in Fig. 9. The order of the two image groups was shuffled in each question to eliminate the bias of choosing the left (or right) group more frequently than the other. Participants were paid $1.00 per HIT.

A.2.2 Data Shown to Participants. The nonwords used in this survey were taken from a dataset that provides 270 randomly-created English nonwords (both spelling and pronunciation) annotated with evoked emotion labels [24]. This dataset was selected to measure the performance of the proposed nonword-to-image generation on totally random nonwords that do not necessarily have specific phonetically similar words. We did not use the nonwords created in Section 4 because their pronunciations were supposed to be quite close to certain existing words, which in most cases could give little difference in the generated images between the proposed and comparative methods.

The audio data for each of the pronunciation of nonwords were prepared using Speech Application Programming Interface (SAPI) [4] on Microsoft Windows 10. It was used to convert the pronunciations of nonwords written with International Phonetic Alphabet (IPA) symbols (e.g., /ˈblaʊəɹ/) into spoken sounds (e.g., audio of an English speaker pronouncing the pronunciation /ˈblaʊəɹ/). Zira (American English, Female) was selected as the speaker with a fixed speaking rate of 0.

[4]https://docs.microsoft.com/en-us/previous-versions/windows/desktop/ee125663(v=vs.85) (Accessed July 11, 2023)

Table 3: Consonant-spelling correspondences used for the quantitative evaluation.

Consonant	/z/	/s/	/v/	/f/	/b/	/p/	/d/	/t/	/g/	/k/		
Spelling	'z'	's'	'v'	'f'	'b'	'p'	'd'	't'	'g'	'k' or 'c'		
Consonant	/ð/	/θ/	/ʃ/	/h/	/m/	/n/	/ɹ/	/j/	/l/	/w/		
Spelling	'th'	'th'	'sh'	'h'	'm'	'n'	'r'	'y'	'l'	'w'		

Table 4: Vowel-spelling correspondences used for the quantitative evaluation.

Free Vowel	/eɪ/	/ii/	/aɪ/	/oʊ/	/uu/	/aʊ/	/ɔɪ/	/ɑɑ/	/ɔ/
Spelling	'a'	'ea'	'i'	'o'	'u'	'ou'	'oi'	'au'	'o'
Checked Vowel	/ʌ/	/æ/	/ɛ/	/ɪ/	/ʊ/				
Spelling	'u'	'a'	'e'	'i'	'u'				

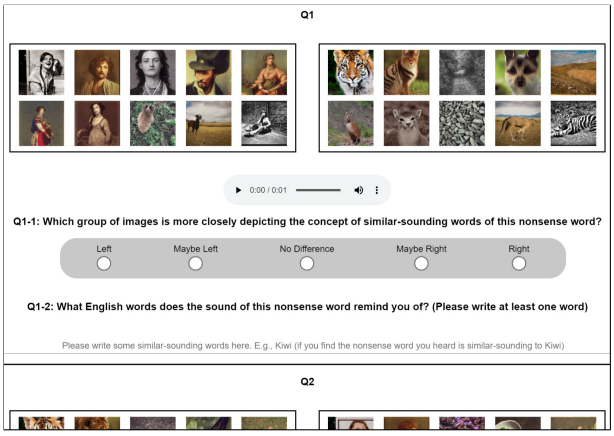


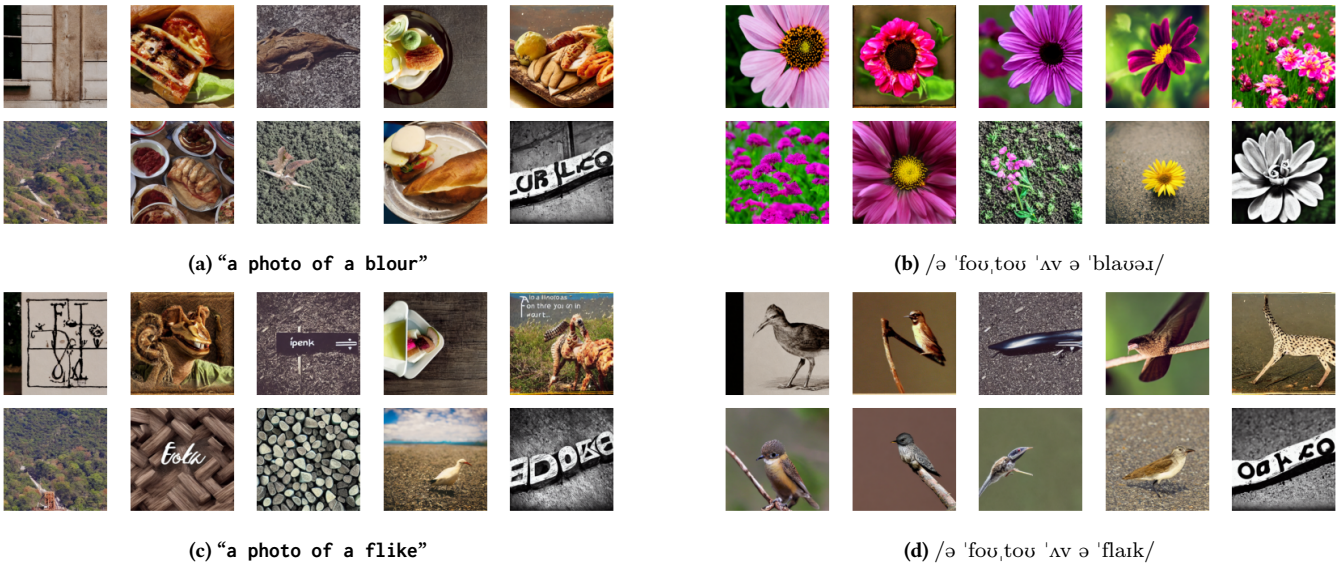
Figure 9: Screenshot of the user interface presented to participants in the qualitative experiment.

A.3 Training German Model

This section describes the detail of the training data used to extend the proposed method to the German language in Section 6.

The training data were constructed from 40,146 pairs of English-German sentences that previous work [1] had created from English using Amazon Translate[5]. The conversion of German words into International Phonetic Alphabet (IPA) transcription was performed using the Wiktionary-based German pronunciation dictionary[6]. Using the entry words of this dictionary as a German wordlist, we also added sentences in the shape of /<WORD>/, /ˈaɪn ˈfooto ˈfɔn <WORD>/ ("Ein Foto von <WORD>"), and /ˈaɪn ˈfooto ˈfɔn ˈaɪnəm <WORD>/ ("Ein Foto von einem <WORD>"), to the training data. We translated those added German sentences into English using Amazon Translate to calculate the CLIP embeddings for them. Regarding the German word-to-pronunciation conversion, because the pronunciation dictionary lacked entries for many of the German

[5]https://aws.amazon.com/translate/ (Accessed July 11, 2023)
[6]https://github.com/Kyubyong/pron_dictionaries/ (Accessed July 11, 2023)



(a) "a photo of a blour"

(b) /ə ˈfoʊˌtoʊ ˈʌv ə ˈblaʊəɹ/

(c) "a photo of a flike"

(d) /ə ˈfoʊˌtoʊ ˈʌv ə ˈflaɪk/

Figure 10: Examples of groups of ten images generated for prompts containing randomly created nonwords [24] used in our qualitative evaluation. The left image groups were generated from texts using an existing T2I generation method Stable Diffusion [31], while the right groups were generated from pronunciations using the proposed P2I generation method.

words in inflected forms like "*Lustige*", which comprises the stem "*Lustig*" and the affix "*-e*", we also trained a Transformer-based grapheme-to-phoneme converter named DeepPhonemizer[7]. We used this converter only when the dictionary had no entry for the word. These operations resulted in the training data of 133,914 pronunciations of German sentences with English translations.

The settings and parameters for training IPA-CLIP were identical to Section 4.3.1 except for the training epochs. We trained the German model up to 600 epochs due to the small size of training data compared to the English one (1,192,804 sentences).

A.4 Preprocessing of Pronunciation Data

Throughout the paper, to precisely calculate phonetic similarity and compare pronunciations among different languages, some modifications were made to the raw pronunciation data that existing dictionaries of each language provide. Specifically, the following six modifications were made to each of the pronunciations:

(1) Primary stress /ˈ/ is inserted at the beginning of pronunciations of monosyllabic words.
(2) Symbol /ː/, which represents the lengthened vowel, is replaced with the previously occurred vowel (/ˈkjuː/ ("*C**ue***") changed to /ˈkjuu/).
(3) Affricates are split into two separate phonemes. Specifically, the four affricates that appear either in English or German, /ʧ/ (as in English "***Ch**oke*" (/ˈʧoʊk/)), /ʤ/ (English "***J**oke*" (/ˈʤoʊk/)), /ʦ/ (German "***Z**eit*" (/ˈʦaɪt/)), and /p͡f/ (German "***Pf**eil*" (/ˈp͡faɪl/)), are split into a combination of two phonemes /tʃ/, /dʒ/, /ts/, and /pf/, respectively.
(4) English rhotic vowels, /ɚ/ (as in "*Mirr**or***": /ˈmɪɹɚ/) and /ɝ/ (as in "***Pearl***": /ˈpɝl/), are both converted to /əɹ/ (e.g., /ˈmɪɹəɹ/).
(5) Allophones are converted to their corresponding phonemes or phones. English /ɾ/ (as in English "*Wa**t**er*" (/ˈwɔɾɚ/)) is replaced with /t/. German /ʁ/ (German "***R**uf*" (/ˈʁuuf/)) and /χ/ (German "*Do**ch***" (/ˈdɔχ/)) are replaced with /ʁ/ and /x/, respectively.
(6) German syllabic consonant /n̩/ (as in "*Trag**en***" (/tʁaagn̩/)) is converted to /ən/.

A.5 Examples of Nonword-to-Image Generation

This section provides some examples of nonword-to-image generation using either the comparative Text-to-Image (T2I) generation method [31] or the proposed Pronunciation-to-Image (P2I) generation method. Figure 10 displays examples of generated images for nonwords used in the qualitative evaluation in Section 5 which were taken from an existing dataset [24].

Received 20 July 2023; accepted 10 August 2023

[7] https://github.com/as-ideas/DeepPhonemizer/ (Accessed July 11, 2023)