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:

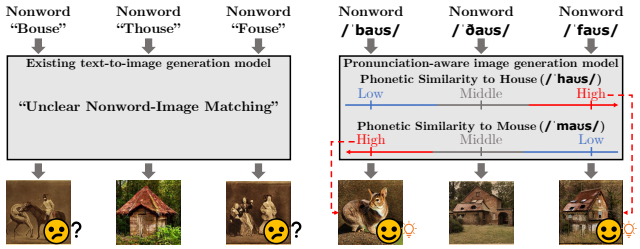
Chihaya Matsuhira, Marc A. Kastner, Takahiro Komamizu, Takatsugu Hirayama, Keisuke Doman, and Ichiro Ide. 2023. Nonword-to-Image Generation Considering Perceptual Association of Phonetically Similar Words. In *Proceedings of the 1st International Workshop on Multimedia Content Generation and Evaluation: 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

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

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

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 /"haʊs/. If a user expects the nonword "*Fouse*" to rhyme with "*House*", it will be transcribed as /"faʊs/. The effect of phonetic similarity on the proposed image generation is illustrated in Fig. 1(b). Given a nonword /"faʊs/, the proposed method generates images of its phonetically similar word "*House*" (/"haʊs/) rather than "*Mouse*" (/"maʊs/). Meanwhile, given another nonword /"baʊs/, 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 Di usion [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 speci c 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 Uccelli , French Oiseaux , German Vögel , and Spanish Pájaros). Such a concatenation of subwords of foreign words meaning a speci c concept can easily produce nonwords that can fool existing T2I generation models [18].

This paper controls Stable Di usion by inserting conditioning pronunciation vectors computed based on phonetic similarity into a pretrained Stable Di usion. 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 Di usion is an open-source latent di usion model for Text-to-Image (T2I) generation [28], which was trained on a huge number of image-text pairs in Large-scale Arti cial 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 de ned 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 Di usion on the pronunciation input, we substitute the conditioning vectors of the CLIP text encoder used in Stable Di usion with the pronunciation-based ones computed by IPA-CLIP. This approach requires much less computational costs and resources than retraining Stable Di usion 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 /@ "foU•toU "2v @ "kaEt/ (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 /"tAAlf@n/ 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/"faUs/ (pronunciation of a nonword Fouse), its pronunciation encoder would compute a phonetic embedding vector similar to the one for House(/"haUs/) rather than Mouse(/"maUs/), 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/"faUs/ 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/"faUs/ is phonetically similar to both House and Souse(/"saUs/), 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 Di usion

The text condition of Stable Di usion requires an $_1$-dimensional last-hidden-state output of the Transformer in the CLIP text encoder (is the maximum length of tokens of the Transformer), while the embedding vector calculated by IPA-CLIP is a $_2$-dimensional nal output of the CLIP text encoder. In the CLIP architecture, the $_2$-dimensional output is computed from the $_1$-dimensional last-hidden-state output. Hence, to insert the embedding vectors of IPA-CLIP into Stable Di usion, we need to perform this operation inversely and reconstruct $_1$-dimensional vectors from $_2$-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 $_2$-dimensional nal output and an $_1$-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 Di usion along with a noise vector to synthesize an image.

4 QUANTITATIVE EVALUATION

To quantitatively evaluate the in uence 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 classi cation dataset [1]. Note that the term Nonword hereafter will denote the pronunciation of a non-existing word, and thus not the spelling (e.g., a nonword/"faUs/).

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./"tAAlf@n/in Fig. 3) created by slightly modifying the pronunciation of a certain existing word (e.g.,Dolphin (/"dAAlf@n/)), the goal is to nd 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/"faUs/ originated from an existing word Mouse(/"maUs/), it should not generate images of a mouse for the nonword since there exists a more phonetically similar word House(/"haUs/). 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 ve classes each. Hence, each nonword has both a ground-truth class and a superclass derived from the original word. For instance, the nonword/"tAAlf@n/ is assigned with the ground-truth class Dolphin and also its superclass Aquatic Mammals Note that the two superclasses identically named Vehiclesare 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 classi cation 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/"tAAlf@n/ are predicted to contain dolphins and are classi ed as Dolphin, the majority class for/"tAAlf@n/ is regarded as the class Dolphin. At the same time, we also calculate the majority superclass for/"tAAlf@n/as the most frequent