

---

# Text-Free Image-to-Speech Synthesis Using Learned Segmental Units

---

<b>Wei-Ning Hsu</b> <sup>*†</sup> MIT wnhsu@csail.mit.edu	<b>David Harwath</b> <sup>*</sup> UT Austin harwath@utexas.edu	<b>Christopher Song</b> JHU csong23@mit.edu	<b>James Glass</b> MIT glass@mit.edu
---	--	---	--

## Abstract

In this paper we present the first model for directly synthesizing fluent, natural-sounding spoken audio captions for images that does not require natural language text as an intermediate representation or source of supervision. Instead, we connect the image captioning module and the speech synthesis module with a set of discrete, sub-word speech units that are discovered with a self-supervised visual grounding task. We conduct experiments on the Flickr8k spoken caption dataset in addition to a novel corpus of spoken audio captions collected for the popular MSCOCO dataset, demonstrating that our generated captions also capture diverse visual semantics of the images they describe. We investigate several different intermediate speech representations, and empirically find that the representation must satisfy several important properties in order to work well.

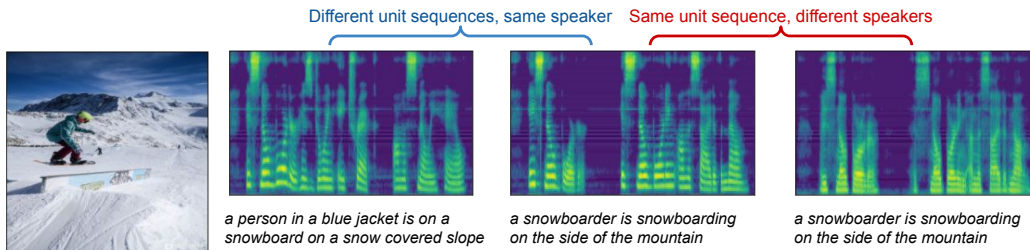


Figure 1: Spoken image captions generated from the proposed model, with diversity in both linguistic content and acoustic properties, controlled through the I2U and the U2S models, respectively. A speech recognition model is used to transcribe speech into text for illustration. Audio samples are available at <https://wnhsu.github.io/image-to-speech-demo>.

## 1 Introduction

Current state-of-the-art models for automatic speech recognition and speech synthesis are based upon supervised machine learning techniques, where the supervision most often takes the form of utterance-level text transcriptions of speech audio. These models are very data-hungry, and collecting transcribed speech audio poses a significant barrier for deploying speech technologies to all the world’s languages. Although there are over 7,000 languages spoken worldwide [64], only several dozen have enough data available to support supervised speech recognition, and many languages do not even employ a writing system [2]. In contrast, most people learn to use spoken language long before they learn to read and write, suggesting that linguistic annotation is not a prerequisite for

---

<sup>\*</sup>Equal contribution

<sup>†</sup>The author performed the work while at MIT, and is now at Facebook AI Research

speech processing systems. This line of reasoning motivates research that aims to discover meaningful linguistic abstractions (phones, words, etc.) directly from the speech signal, with the intention that they could reduce the reliance of spoken language systems on text transcripts.

A rich body of work has recently emerged investigating representation learning for speech using visual grounding objectives [92, 32, 37, 57, 39, 74, 15, 4, 85, 47, 58, 91, 54, 25], as well as how word-like and subword-like linguistic units can be made to emerge within these models [33, 36, 23, 4, 36, 34, 40, 35]. So far, these efforts have predominantly focused on *inference*, where the goal is to learn a mapping from speech waveforms to a semantic embedding space. *Generation* of speech conditioned on a point in a semantic space has been less explored, and is what we focus on in this work. We hypothesize that generative approaches offer interesting advantages over relying solely on inference. For example, prior works have demonstrated the capability of recognizing visually descriptive words, but have not been shown to learn non-visual words or grammar. Our experiments show that these aspects of spoken language are learned to some degree by a visually-grounded generative model of speech.

Specifically, we introduce a model capable of directly generating fluent spoken audio captions of images without the need for natural language text, either as an intermediate representation or a form of supervision during training (Figure 1). Tremendous progress has been made recently in natural language image caption generation [62, 22, 73, 102, 59, 109, 68, 84, 17, 70, 6, 71, 114, 18] and naturalistic text-to-speech synthesis [104, 97, 79, 93, 105, 88, 75, 81]. Combining these models provides a means for generating spoken image descriptions, but existing approaches for training these models are reliant on text during training. Instead, we leverage sub-word speech units discovered using a self-supervised learning objective as a drop-in replacement for the text. We hypothesize that by using such techniques, an even wider variety of traditionally text-based NLP models could be applied to speech data without the need for transcription or ASR. Because all human languages utilize small, discrete phonetic inventories, we posit that our framework should be applicable for any language in the world. In our experiments, we demonstrate that not just any set of discovered speech units can function in this role. We find the greatest success with units that are *discrete*, exhibit a *low frame-rate*, and *highly robust* to speaker and environmental variability.

The main contributions of our paper are as follows:

- 1. The first methodology for fluent image-to-speech synthesis that does not rely on text.** A critical aspect of our approach is factorizing the model into an Image-to-Unit (I2U) module and a Unit-to-Speech (U2S) module, where the speech units are discovered in a self-supervised fashion. This approach enables disentanglement of linguistic variability and acoustic/speaker variability.
- 2. Extensive analysis on the properties required for learned units to replace text.** While the idea may seem simple and straightforward, it is not a trivial task to obtain proper units. In fact, most of the units experimented in this paper fail to serve as drop-in replacements. Moreover, we demonstrate that what are deemed good units vary significantly for inference and generation, respectively.
- 3. Proposing a semantic diversity-aware captioning metric for sampling-based evaluation.** We show that even when an I2U model fails to generate sensible caption through beam search decoding, it can still produce reasonable captions by sampling from the posterior, hinting that posterior mode-based evaluation can only inspect limited aspects of a model. Hence, we advocate for sampling-based evaluation and propose a new metric to assess diversity.
- 4. Over 600,000 Spoken audio captions for the MSCOCO dataset.** We collect speech from 2,352 people who were tasked with reading each MSCOCO caption out loud. This dataset (SpokenCOCO) is made publicly available at <https://groups.csail.mit.edu/sls/downloads/spokencoco/> to support work at the intersection of speech, language, and vision.

## 2 Related Work

**Image-to-Text and Image-to-Speech Captioning.** Significant progress towards generating realistic (text) captions that describe the content of visual images was made with the advent of deep neural networks [102, 59, 109, 6]. Far less work has focused on generating spoken audio captions from natural images. Training an image-to-speech system using separate (*image, text*) and (*text, speech*) datasets was explored in [72]. One other work has explored image-to-speech synthesis without using text [38], but with limited results. In [38], BLEU scores were only computed in terms of unsupervised acoustic units, not an estimate of the actual words produced by the synthesizer. The resulting captions

were not evaluated for fluency, naturalness, or intelligibility, and the BLEU scores in terms of the unsupervised units were very low (0.014 on the MSCOCO test set).

**Speech Synthesis without Text and Voice Conversion.** Voice conversion is a classic problem in speech processing that involves resynthesizing a recording to sound as if it was spoken by a different person [1, 90, 95]. Voice conversion has recently seen progress using neural approaches [45, 49, 50, 46, 27, 12, 14, 69, 56, 87], but the most relevant work to our own is the ZeroSpeech 2019 challenge [24, 94, 10], which explicitly addresses the unsupervised learning of discrete speech units which can be used as the basis of a synthesizer.

**Speech Representation for Downstream Tasks.** Pre-training models on ImageNet [19] is now par for the course in computer vision, and NLP has also recently embraced the pre-training paradigm [78, 21, 111]. Pre-training is appealing because it enables powerful models with many parameters to be fit to a large amount of domain-general training data, and then fine-tuned on small amounts of domain-specific data relevant for a particular downstream task. The speech community has not yet discovered a pre-training approach whose performance and wide applicability is comparable with the state-of-the-art techniques in vision or NLP, but many recent works have been attempting to do just that. Several papers have investigated masked prediction [8, 86, 7, 103], while others have implemented autoencoder models with various constraints applied to the latent representations, such as discreteness [98, 25, 11] or disentanglement [50, 48, 65, 60]. Predictive coding has a long history in speech processing, but has recently seen a revival using deep neural networks [76, 16]. Most relevant to our work is the use of multimodal grounding to learn discrete speech representations [35].

### 3 Learning to Speak By Only Seeing and Listening

#### 3.1 Overview of the Proposed Framework

A depiction of our modeling approach is shown in Figure 2. Caption generation for an image involves a cascade of two components: given an input image  $I$ , we first generate a linguistic unit sequence  $U$  according to the I2U module  $P(U | I)$ . Given the linguistic symbol sequence  $U$ , we generate a speech waveform  $S$  according to the U2S module  $P(S | U)$ . If the linguistic unit sequence  $U$  were to take the form of natural language text, the model would be equivalent to the cascade of a conventional image captioning system followed by a TTS module. Note that we assume  $S \perp I | U$  because prosody variation is not dependent on the image for the datasets considered.

The key idea in this paper is to instead define  $U$  to be a sequence of *learned* speech units, discovered without text supervision. Specifically, we use the visually-grounded sub-word speech unit discovery model introduced by [35]. We define inference with this model as  $U = f(S)$ , enabling us to “transcribe” any given speech audio waveform  $S$  into a sequence of units  $U$ . The addition of this third component enables us to train  $P(U | I)$  from a dataset of images paired with spoken captions  $\{(I_1, S_1), \dots, (I_N, S_N)\}$ . The conditional independence assumption between  $S$  and  $I$  given the  $U$  enables us to choose any arbitrary speech dataset for training  $P(S | U)$ , therefore enabling the speaker characteristics and other acoustic properties to be independently controllable from the I2U system [106, 51, 89, 42, 3, 112, 30].

#### 3.2 Datasets

For training the  $P(U | I)$  model as well as the sub-word speech unit model  $f(S)$ , we rely on a dataset of images and their accompanying spoken audio captions. We consider 3 datasets of this type: **SpokenCOCO** is based on the MSCOCO captioning dataset [67], but as part of this work we collect a spoken version of the 605,495 captions via Amazon Mechanical Turk by displaying the text to a person and having them read it aloud. The captions were recorded by 2,352 different speakers. We provide additional details regarding the dataset statistics in the supplementary material. **Flickr8k Audio Captions** [32] contains 8,000 images and 40,000 text captions from the Flickr8k dataset [83], which have been read aloud by 183 different speakers. **Places Audio** [37] is comprised of 400,000 images from the MIT Places 205 dataset [113] that have each been captioned by a single person. Unlike SpokenCOCO and the Flickr8k audio captions, the Places Audio captions are spontaneous speech and were collected from over 2,500 different speakers.

While in principle we could also use the speech from any of the captioning datasets to train the symbol-to-speech model  $P(S | U)$ , in practice the naturalness and intelligibility of speech produced

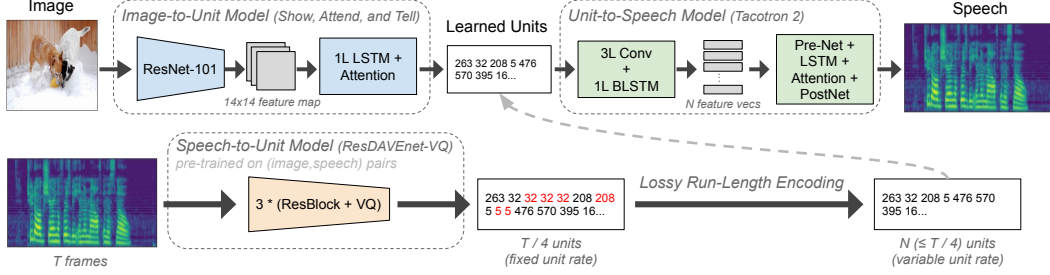


Figure 2: Diagram of our proposed framework.

by a synthesis model benefits from training data recorded by a single speaker in a clean recording environment [53, 110, 29, 52]. For this reason, we choose to use the waveforms from two well-known speech synthesis training datasets: **LJSpeech** [55] is a dataset of approximately 24 hours of speech drawn from audio books, recorded by a single speaker. **VCTK** [99] is a dataset often used for multi-speaker synthesis and contains speech recordings from 109 different speakers, each of whom read aloud 400 sentences sampled from newspapers. Note that using different datasets for training each module constructs a very challenging setup, where the learned units need to contain sufficient linguistic information and minimal acoustic information for the entire pipeline to function properly.

### 3.3 Learning Robust Linguistic Units from Visually-Grounded Speech

Before training the I2U or U2S modules, we must train the transcription model  $f(S)$ . To do this, we turn to the ResDAVENet-VQ model introduced by [35]. This model learns a mapping from spoken caption waveforms and visual images to points in a shared semantic space, in which inner products are predictive of semantic relatedness. The intermediate layers of the ResDAVENet-VQ model incorporate vector quantization (VQ) blocks, which were shown in [35] to learn phone-like and word-like tokenizations of an input speech waveform. We investigate both the VQ2 (higher framerate) and VQ3 (lower framerate) discrete unit representations from a ResDAVENet-VQ model that was trained using a  $\{2\} \rightarrow \{2, 3\}$  curriculum (in the notation given in [35]).

The majority of un/self-supervised speech representations are trained with a signal reconstruction-based objective [11, 50] or predictive coding [76, 16]. In contrast, ResDAVENet-VQ and related models of visually grounded speech are trained to associate speech with contextually relevant visual inputs using a contrastive loss such as the triplet loss [107, 37] or noise contrastive estimation (NCE) [76, 54], which maximizes a lower bound on the mutual information between  $I$  and  $S$  as observed in [96]. This objective forces the learned speech representations to be useful for predicting image semantics. Because visual semantics are described with words, the units learned by ResDAVENet-VQ are forced to be predictive of words rather than speaker identity, acoustic environment, noise, etc., all of which must be modeled by reconstruction-based approaches.

Although the ResDAVENet-VQ model has been shown to be capable of learning a hierarchy of discrete speech units from phone-like to word-like, the experiments in [35] show that only several hundred words are explicitly learned by the model (a small subset different words appearing in the Places Audio captions), which tend to be “visual words.” Conversely, the phone-like units learned by the lower VQ layers of the model were shown to cover all of the phones in American English (as there are only several dozen). For this reason, to represent  $U$  in our I2U model, we choose to use the representations learned by the lower VQ layers that were shown to be correlated with phones. We postulate that future work may benefit from incorporating the word-level ResDAVENet-VQ units into our framework. Once the ResDAVENet-VQ model has been trained, we can use it to “transcribe” any given speech waveform into a sequence of units by performing a forward pass of the speech through the audio branch of the model, retaining only the outputs of the target VQ layer(s).

Nominally, the VQ layers will output codebook vectors (where each unique vector represents a speech unit) at a uniform temporal rate, which is downsampled with respect to the framerate of the acoustic input depending upon which VQ layer is used. Given an input spectrogram to the model computed with a 10ms frame shift, the two VQ layers we investigate in this paper (VQ2 and VQ3) respectively output vectors every 20ms and 40ms. In general, the VQ units from each model are repeated for several consecutive frames. We can decrease the average length of the symbol sequence

$U$  by employing a lossy form of run-length encoding (RLE) (see Figure 2) which retains the sequence of symbol identities but discards duration information. This removes the burden of unit duration modeling from the I2U model and shifts it onto the U2S model, which we will show to be crucial.

### 3.4 Image-to-Unit and Unit-to-Speech Models

Both the I2U model and the U2S model are based upon recurrent seq2seq with attention networks [9]. Specifically, we use Show-Attend-and-Tell [109] for the I2U model. We use the off-the-shelf implementation that has an image encoder pre-trained for classification (which is language agnostic and hence should work in any language within our proposed framework). We choose this model not only because of its ubiquity, but also because it represents one of the simplest captioning models by today’s standards. After inferring the discrete unit sequence  $U$  for each speech recording  $S$  in the training dataset using the ResDAVENet-VQ model, we train the I2U model on the resulting  $(I, U)$  pairs of images and their speech unit sequence representing the caption. For the U2S model, we use the off-the-shelf implementation of Tacotron-2 [88] with no pre-training to generate spectrograms, and WaveGlow [81] to generate waveforms from these spectrograms. TTS models generally require training pairs of the form  $(text, speech)$ , but in our case the ResDAVENet-VQ unit transcription model enables us to train the Tacotron-2 model on any arbitrary set of speech waveforms by transcribing each recording  $S$  into its unit sequence  $U$ , and then training the model on the  $(U, S)$  pairs. Both models are trained with the maximum likelihood objective, with the I2U model trained to maximize  $\mathbb{E}_{I,U} [\log P(U | I)]$  and the U2S model trained to maximize  $\mathbb{E}_{S,U} [\log P(S | U)]$ . This has interesting implications for the diversity of captions produced with different decoding strategies, which we explore in Section 4.

## 4 Experiments

Our experiments attempt to address several questions:

**First, how can we measure the performance of an image-to-speech system?** Generating a spoken caption for an image within our framework first involves generating a latent unit sequence with the I2U model, followed by synthesizing a speech waveform conditioned on the unit sequence with the U2S model. The concatenation of these models can fail to produce a good caption if the I2U system fails to encode linguistic/semantic information into the unit sequence, or if the U2S system fails to synthesize an intelligible waveform given a unit sequence. To better localize these failure modes, we evaluate the full I2S system as well as the U2S system in isolation. We evaluate the U2S system by using it as a vocoder and soliciting human judgements in the form of Mean Opinion Score (MOS) and Side-By-Side (SXS) preference tests. To evaluate the full I2S system, we could use any method that measures the semantic information contained in the generated language. One way would be to use a speech-to-image retrieval system, while another is to infer a text transcription of the generated speech with an ASR system, and then employ standard text-based caption evaluation metrics. This has the advantages of making possible a comparison with a "upperbound" system in the form of text-based captioning, as well as measuring other aspects of language generation such as syntactic correctness (via BLEU-4) and scope of the learned vocabulary. For those reasons, we use the latter strategy.

**Second, what properties must learned units have in order to be a drop-in replacement for text?**

The most essential differences between text and speech are the amount of information encoded and the sequence lengths. Beyond text, speech also encodes prosody, speaker, environment information as well as the duration for each linguistic unit, all of which are minimally correlated with the conditioned images. We hypothesize that learned speech units should discard such information in order to seamlessly connect the I2U and U2S modules. To verify it, we pay particular attention to the variations of the learned units in *frame rate* (VQ2 or VQ3), *encoding of duration information* (using RLE or not), and *robustness to domain shift* (WaveNet-VQ or ResDAVENet-VQ).

**Third, how should language generation models be evaluated more generally?** We examine evaluation of the I2S model using beam search-based decoding as well as sampling-based decoding. We find that because evaluation metrics that are reliant on beam search-based decoding only evaluate the mode of a model’s posterior, they do not reflect the ability of a model to generate diverse linguistic content. Furthermore, we show that it is possible for a model’s posterior mode to be linguistically meaningless, and yet meaningful language can still be generated with sampling-based decoding. Towards this end, we introduce a novel multi-hypothesis evaluation metric (M-SPICE).

#### 4.1 Evaluating the Unit-to-Speech Model

We first evaluate the U2S model on several different intermediate unit representations. We consider the VQ2 and VQ3 units derived from the ResDAVENet-VQ model described in Section 3.3, as well as units derived from the unimodal WaveNet-VQ model [11, 10] trained on the same 400,000 Places Audio captions as the ResDAVENet-VQ model. We construct a Tacotron-2 model for each of the three unit types on the LJSpeech audio data. We transcribe each LJSpeech utterances into a unit sequence, then train the U2S model from the resulting unit sequence and waveform pairs. We evaluate the naturalness of the speech produced by each model on held-out data, both in-domain using LJSpeech and out-of-domain using SpokenCOCO. Amazon Mechanical Turk (AMT) workers performed Side-by-Side preference tests (SXS) and naturalness evaluation based on mean opinion scores (MOS) on a scale from 1 to 5 for each U2S model, which we display in Table 1. Although VQ2 was preferred for in-domain synthesis on LJSpeech, VQ3 achieved the highest scores on the out-of-domain task (SpokenCOCO), indicating that of the three unit types it has the strongest robustness to domain shift.

Table 1: (left) Subjective preference (%) between unit representations. (right) Subjective evaluation on fidelity (MOS with 95% confidence interval).

Unit A	Unit B	LJSpeech			SpokenCOCO			Unit	LJSpeech MOS	SpokenCOCO MOS
		A	Neutral	B	A	Neutral	B			
VQ3	VQ2	23.9	31.5	44.6	40.4	32.5	27.5	VQ3	$3.723 \pm 0.039$	$3.336 \pm 0.044$
VQ3	WVQ	36.6	37.1	26.3	58.3	21.8	19.9	VQ2	$3.932 \pm 0.036$	$2.961 \pm 0.045$
VQ2	WVQ	43.9	30.2	25.9	49.8	24.2	26.0	WVQ	$3.658 \pm 0.040$	$2.896 \pm 0.053$

#### 4.2 Incorporating the Image-to-Unit Model

We trained a Show, Attend, and Tell (SAT) model on SpokenCOCO for the WaveNet-VQ, VQ2, and VQ3 units, as well as VQ3 units without RLE. We also compare to text characters and words; the full hyperparameter and training details for all models are provided in the supplementary material, but in general we kept these as constant as possible when comparing different linguistic representations.

Before connecting the U2S model, we noticed that all speech unit models except the one trained on RLE VQ3 units failed during *beam search* decoding on the test images; rather than producing a diverse sequence of output units, the decoder would generally get stuck in a loop until the maximum decoding length was reached. This phenomenon also took place using VQ3 units without RLE, indicating that the decoder had difficulty modeling the duration of the units. We hypothesize that the reason the VQ2 and WaveNet-VQ units failed is due to their lack of invariance to domain shift, as evidenced by their decay in naturalness when used for out-of-domain synthesis as shown in Table 1. For this reason, the main results in this section using beam search are only shown for the VQ3 model. We show example outputs of each system in the supplementary material, as well as BLEU-4 scores on SpokenCOCO in terms of the speech units. The looping phenomenon in the context of beam search decoding with generative models of language was also observed by [44, 26, 82, 43, 63, 108]; we explore sampling-based alternatives to beam search for the other unit types later in Section 4.3.

To evaluate the full Image-to-Speech model, we first train an automatic speech recognition (ASR) system on the re-synthesized SpokenCOCO captions using the VQ3 Tacotron-2 model. Specifically, we use the Wav2Letter++ [80] implementation of the seq2seq+TDS model [31], using the COCO caption text as the ground truth labels for ASR training. This enables us to estimate a word-level transcription of the spoken captions produced by our system, which we can then use to compute standard captioning evaluation metrics (including BLEU-4 [77], METEOR [20], ROUGE-L [66], CIDEr [100], and SPICE [5]) and make direct comparisons with conventional text captioning systems. In order to verify that the synthesized captions are intelligible to humans and the ASR system did not simply learn to recognize artifacts of the synthesized speech, we asked AMT workers to transcribe into words a set of 500 captions generated by our system and also evaluated their naturalness. 3 workers transcribed and 3 workers rated each caption, allowing us to compute an MOS score ( $3.615 \pm 0.038$ ), a word error rate (WER) between the 3 human transcriptions (9.40%), as well as between the human and the ASR-produced transcriptions (13.97%) averaged against all 3 transcriptions.

Table 2a summarizes our results. We compare with recent approaches from the literature for bottom-up text captioning (top 3 rows). We train the decoder of an SAT model while keeping the image encoder fixed (middle 3 rows), in addition to fine-tuning the image encoder (bottom 3 rows). We

report results when training the I2U model to output words, characters, or VQ3 units, both with greedy decoding and beam search decoding. Despite having no access to text, the SAT-FT speech captioning model trained on VQ3 units achieves a BLEU-4 score of .233 with beam search decoding. This is very close to the .243 achieved by the original SAT word-based captioning model. Figure 1 shows that the generated captions are fluent and reflect the implicit learning of some syntactic rules. We also report captioning scores on the smaller Flickr8k spoken caption dataset in Table 2b.

Table 2: Word-based caption evaluation using **BLEU-4**, **METEOR**, **ROUGE**, **CIDEr**, and **SPICE**. ASR is used to transcribe the spoken captions generated by the proposed VQ3 model into text for evaluation. The beam size  $\in \{3, 5, 8, 10\}$  was chosen for each model to maximize SPICE. Our word-based SAT models outperform [109] because we use a stronger image encoder (ResNet-101).

model	$U$	B-4	M	R	C	S
SAT [109]	word	0.243	0.239	-	-	-
AA [70]	word	0.327	0.260	0.540	1.042	-
SAT	word	0.315	0.253	0.533	0.984	0.185
	char	0.289	0.239	0.512	0.879	0.172
	VQ3	0.186	0.186	0.446	0.584	0.127
SAT-FT	word	0.339	0.265	0.551	1.062	0.196
	char	0.323	0.256	0.536	1.002	0.187
	VQ3	0.233	0.212	0.478	0.732	0.149

model	$U$	B-4	M	R	C	S
SAT [109]	word	0.213	0.203	-	-	-
SAT	word	0.216	0.207	0.469	0.550	0.149
	char	0.190	0.190	0.441	0.476	0.136
	VQ3	0.116	0.141	0.390	0.232	0.091
SAT-FT	word	0.225	0.215	0.483	0.584	0.155
	char	0.191	0.196	0.450	0.519	0.143
	VQ3	0.125	0.145	0.391	0.245	0.095

### 4.3 From Mode to Distribution: Sampling-Based Caption Evaluation

The results in the previous section only evaluate beam search decoding with the I2U model, and do not fully reveal the posterior over captions for an input image, or whether the unit representations that failed with beam search would work well with other methods. To probe this, we evaluate the models using sampling-based caption generation. Figure 3 shows the SPICE scores on SpokenCOCO using beam search and two sampling-based methods. VQ3 still performs the best of all unit types with both beam search and sampled decoding. VQ3 can sometimes generate captions with beam search when the beam is kept small, but as the beam grows it begins to loop and the scores become very low. We see that all unit types can generate reasonable captions when using sampling-based decoding, although ResDAVENet-VQ units consistently outperform the WaveNet-VQ units, suggesting that they better capture sub-word structure. Example outputs are provided in the supplementary material.

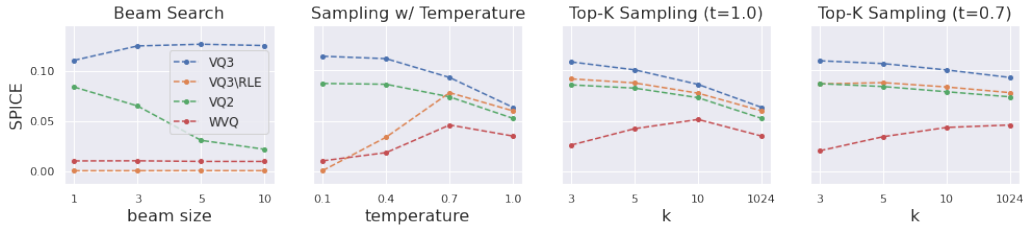


Figure 3: MSCOCO test SPICE scores of various units and decoding methods. VQ3/RLE denotes VQ3 units without RLE. Top- $k$  sampling considers only the  $k$ -most probable units at each step.

We estimated the vocabulary size of our SAT-FT model using VQ3 units by counting the number of unique recognized words produced at least 3 times when captioning the SpokenCOCO test images. These numbers are shown for the model under the various decoding methods in Figure 4. The number of captions per image is denoted by  $n$ , where top candidates are used for beam search and i.i.d. samples are drawn for sampling. Sampling-based decoding reveals a larger vocabulary size than beam search, and the number of words learned by our models (several thousand) is far greater than the number of words learned by the ResDAVENet-VQ model (approx. 279) in [35]. We hypothesize that training a model to *generate* spoken captions encourages it to learn many more words than only being trained to retrieve images from captions. We also hypothesize that because beam search attempts to find the mode of the posterior distribution over captions, it tends to produce a smaller set of common words and does not reveal the breadth of the distribution learned by the model. To better measure the diversity of our models, we introduce a new captioning metric in the next section.



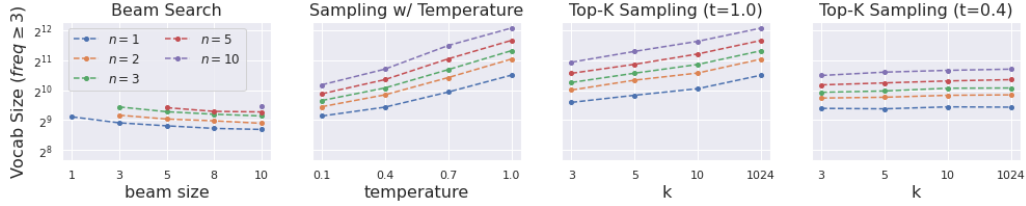


Figure 4: Vocabulary size learned by the proposed image-to-speech model. A word is added to the vocabulary if it appears at least three times from  $n$  runs of decoding on the MSCOCO test set *and* is in a reference dictionary. The same candidates used in Figure 5 are used here for each configuration.

#### 4.4 A New Diversity-Aware Metric: Multi-Hypothesis SPICE

In the previous section, we showed that sampling-based decoding revealed a much larger model vocabulary than beam search, even when the SPICE scores were comparable. This highlights a limitation of SPICE in measuring the *diversity* of the captions. This was previously addressed by [101] who used oracle scoring among the top hypotheses, but in that work each caption was still independently scored. We introduce M-SPICE (Multi-Hypothesis SPICE), which evaluates a *set* of candidate captions against a set of reference captions. To compute M-SPICE for an image, we sample a set of candidate captions (or top  $N$  from beam search) and compute a scene graph  $G(C)$  from the *union of the candidate scene graphs*. We compute the reference scene graph  $G(S)$  using the union of the reference caption scene graphs as is normally done in SPICE. We then evaluate the F-score between  $G(C)$  and  $G(S)$ , which we refer to as M-SPICE. When evaluating sampled captions, we gain a broader estimate of the model’s posterior distribution, rather than just the mode. This addresses the recall penalty faced by single caption evaluation strategies, because the model is considered to have retrieved an object if any of the sampled captions refers to it. Note that simply taking the mean SPICE over the set of samples does not measure the diversity between candidates. Figure 5 shows the M-SPICE scores of our SAT-FT model using VQ3 units on SpokenCOCO. When evaluating over multiple captions ( $n > 1$ ), using the beam search hypotheses increases the score less than sampling.

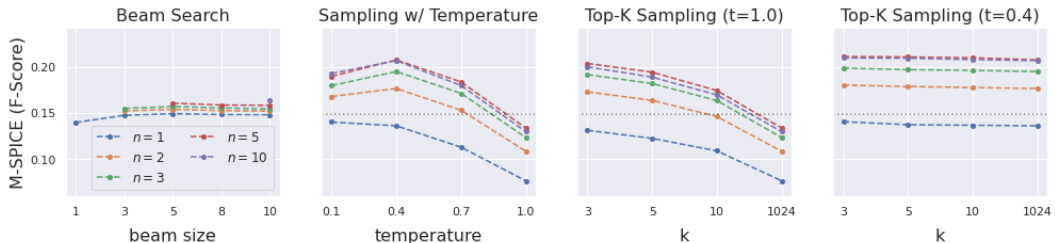


Figure 5: M-SPICE score on the SpokenCOCO test set with different candidate proposal methods. To compute  $n$ -candidate M-SPICE scores, top- $n$  candidates are used for beam search, whereas  $n$  random samples drawn with the given  $(t, k)$  configuration are used for sampling. Black dashed lines show the highest value reported for beam search when  $n = 1$ .

## 5 Conclusion

In this paper, we have presented a novel methodology for Image-to-Speech synthesis without relying on text. We introduced a novel dataset of over 600k spoken captions for the MSCOCO dataset which we make publicly available to the research community. Finally, we introduced a novel captioning metric, M-SPICE, which jointly evaluates the quality and diversity of a captioning model. Future work should investigate applying the proposed method to additional languages, devising improved speech unit representations, and jointly training the speech unit model with the I2S model. This would offer the opportunity to explore new analysis-by-synthesis training objectives. Finally, other tasks such as visual dialog could be explored as a way of teaching the model via richer interaction.



## References

- [1] Masanobu Abe, Satoshi Nakamura, Kiyohiro Shikano, and Hisao Kuwabara. Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan*, 11(2):71–76, 1990.
- [2] Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambourou, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, et al. Breaking the unwritten language barrier: The BULB project. *Procedia Computer Science*, 81:8–14, 2016.
- [3] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Expressive speech synthesis via modeling expressions with variational autoencoder. *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2018.
- [4] Afra Alishahi, Marie Barking, and Grzegorz Chrupała. Encoding of phonology in a recurrent neural model of grounded speech. In *Proc. ACL Conference on Natural Language Learning (CoNLL)*, 2017.
- [5] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Proc. IEEE European Conference on Computer Vision (ECCV)*, 2016.
- [6] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [7] Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*, 2019.
- [8] Alexei Baevski, Steffen Schneider, and Michael Auli. vq-wav2vec: Self-supervised learning of discrete speech representations. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.
- [9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.
- [10] Suhee Cho, Yeonjung Hong, Yookyunk Shin, and Youngsun Cho. VQVAE with speaker adversarial training, 2019.
- [11] Jan Chorowski, Ron J. Weiss, Samy Bengio, and Aaron van den Oord. Unsupervised speech representation learning using wavenet autoencoders. *IEEE Transactions on Audio, Speech and Language Processing*, 2019.
- [12] Jan Chorowski, Ron J Weiss, Rif A Saurous, and Samy Bengio. On using backpropagation for speech texture generation and voice conversion. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.
- [13] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2015.
- [14] Ju-chieh Chou, Cheng-chieh Yeh, Hung-yi Lee, and Lin-shan Lee. Multi-target voice conversion without parallel data by adversarially learning disentangled audio representations. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2018.
- [15] Grzegorz Chrupała, Lieke Gelderloos, and Afra Alishahi. Representations of language in a model of visually grounded speech signal. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.
- [16] Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James R. Glass. An unsupervised autoregressive model for speech representation learning. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.
- [17] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2017.
- [18] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.
- [19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

- [20] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 2014.
- [21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.
- [22] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [23] Jennifer Drexler and James Glass. Analysis of audio-visual features for unsupervised speech recognition. In *Proc. Grounded Language Understanding Workshop*, 2017.
- [24] Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W. Black, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. The zero resource speech challenge 2019: TTS without T. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.
- [25] Ryan Eloff, Herman Engelbrecht, and Herman Kamper. Multimodal one-shot learning of speech and images. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [26] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.
- [27] Fuming Fang, Junichi Yamagishi, Isao Echizen, and Jaime Lorenzo-Trueba. High-quality nonparallel voice conversion based on cycle-consistent adversarial network. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.
- [28] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)*, 32(2):236–243, 1984.
- [29] Nishant Gurunath, Sai Krishna Rallabandi, and Alan Black. Disentangling speech and non-speech components for building robust acoustic models from found data. *arXiv preprint arXiv:1909.11727*, 2019.
- [30] Raza Habib, Soroosh Mariooryad, Matt Shannon, Eric Battenberg, RJ Skerry-Ryan, Daisy Stanton, David Kao, and Tom Bagby. Semi-supervised generative modeling for controllable speech synthesis. *arXiv preprint arXiv:1910.01709*, 2019.
- [31] Awni Hannun, Ann Lee, Qiantong Xu, and Ronan Collobert. Sequence-to-sequence speech recognition with time-depth separable convolutions. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.
- [32] David Harwath and James Glass. Deep multimodal semantic embeddings for speech and images. In *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015.
- [33] David Harwath and James Glass. Learning word-like units from joint audio-visual analysis. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017.
- [34] David Harwath and James Glass. Towards visually grounded sub-word speech unit discovery. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [35] David Harwath, Wei-Ning Hsu, and James Glass. Learning hierarchical discrete linguistic units from visually-grounded speech. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.
- [36] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. *International Journal of Computer Vision*, 2019.
- [37] David Harwath, Antonio Torralba, and James R. Glass. Unsupervised learning of spoken language with visual context. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2016.
- [38] Mark Hasegawa-Johnson, Alan Black, Lucas Ondel, Odette Scharenborg, and Francesco Ciannella. Image2speech: Automatically generating audio descriptions of images. In *International Conference on Natural Language, Signal and Speech Processing*, 2017.
- [39] William Havard, Jean-Pierre Chevrot, and Laurent Besacier. Models of visually grounded speech signal pay attention to nouns: a bilingual experiment on English and Japanese. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

- [40] William Havar, Jean-Pierre Chevrot, and Laurent Besacier. Word recognition, competition, and activation in a model of visually grounded speech. In *Proc. ACL Conference on Natural Language Learning (CoNLL)*, 2019.
- [41] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [42] Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, and Junichi Yamagishi. Deep encoder-decoder models for unsupervised learning of controllable speech synthesis. *arXiv preprint arXiv:1807.11470*, 2018.
- [43] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.
- [44] Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. Learning to write with cooperative discriminators. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.
- [45] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from non-parallel corpora using variational auto-encoder. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2016.
- [46] Chin-Cheng Hsu, Hsin-Te Hwang, Yi-Chiao Wu, Yu Tsao, and Hsin-Min Wang. Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks. *arXiv preprint arXiv:1704.00849*, 2017.
- [47] Wei-Ning Hsu and James Glass. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*, 2018.
- [48] Wei-Ning Hsu and James Glass. Scalable factorized hierarchical variational autoencoder training. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2018.
- [49] Wei-Ning Hsu, Yu Zhang, and James Glass. Learning latent representations for speech generation and transformation. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017.
- [50] Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2017.
- [51] Wei-Ning Hsu, Yu Zhang, Ron Weiss, Heiga Zen, Yonghui Wu, Yuan Cao, and Yuxuan Wang. Hierarchical generative modeling for controllable speech synthesis. In *Proc. International Conference on Learning Representations (ICLR)*, 2019.
- [52] Wei-Ning Hsu, Yu Zhang, Ron J Weiss, Yu-An Chung, Yuxuan Wang, Yonghui Wu, and James Glass. Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [53] Qiong Hu, Erik Marchi, David Winarsky, Yannis Stylianou, Devang Naik, and Sachin Kajarekar. Neural text-to-speech adaptation from low quality public recordings. In *Speech Synthesis Workshop*, volume 10, 2019.
- [54] Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. Large-scale representation learning from visually grounded untranscribed speech. In *Proc. ACL Conference on Natural Language Learning (CoNLL)*, 2019.
- [55] Keith Ito. The LJ speech dataset. <https://keithito.com/LJ-Speech-Dataset/>, 2017.
- [56] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. StarGAN-VC: Non-parallel many-to-many voice conversion using star generative adversarial networks. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2018.
- [57] Herman Kamper, Shane Settle, Gregory Shakhnarovich, and Karen Livescu. Visually grounded learning of keyword prediction from untranscribed speech. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017.
- [58] Herman Kamper, Gregory Shakhnarovich, and Karen Livescu. Semantic speech retrieval with a visually grounded model of untranscribed speech. *IEEE Transactions on Audio, Speech and Language Processing*, PP:1–1, 09 2018.

- [59] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [60] Sameer Khurana, Shafiq Rayhan Joty, Ahmed Ali, and James Glass. A factorial deep markov model for unsupervised disentangled representation learning from speech. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [61] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.
- [62] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In *Proc. International Conference on Machine Learning (ICML)*, 2014.
- [63] Ilya Kulikov, Alexander Miller, Kyunghyun Cho, and Jason Weston. Importance of search and evaluation strategies in neural dialogue modeling. In *Proceedings of the 12th International Conference on Natural Language Generation*, 2019.
- [64] M. Paul Lewis, Gary F. Simon, and Charles D. Fennig. *Ethnologue: Languages of the World, Nineteenth edition*. SIL International. Online version: <http://www.ethnologue.com>, 2016.
- [65] Yingzhen Li and Stephan Mandt. Disentangled sequential autoencoder. In *Proc. International Conference on Machine Learning (ICML)*, 2018.
- [66] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.
- [67] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.
- [68] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [69] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*, 2018.
- [70] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [71] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [72] Shuang Ma, Daniel McDuff, and Yale Song. Unpaired image-to-speech synthesis with multimodal information bottleneck. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019.
- [73] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proc. International Conference on Learning Representations (ICLR)*, 2015.
- [74] Danny Merkx, Stefan L. Frank, and Mirjam Ernestus. Language learning using speech to image retrieval. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.
- [75] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.
- [76] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [77] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.
- [78] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proc. Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.

- [79] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. *arXiv preprint arXiv:1710.07654*, 2017.
- [80] V. Pratap, A. Hannun, Q. Xu, J. Cai, J. Kahn, G. Synnaeve, V. Liptchinsky, and R. Collobert. Wav2letter++: A fast open-source speech recognition system. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [81] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [82] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.
- [83] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using Amazon’s Mechanical Turk. In *Proc. NAACL Conference on Human Language Technologies (NAACL-HLT)*, 2010.
- [84] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [85] Odette Scharenborg, Laurent Besacier, Alan W. Black, Mark Hasegawa-Johnson, Florian Metze, Graham Neubig, Sebastian Stüker, Pierre Godard, Markus Müller, Lucas Ondel, Shruti Palaskar, Philip Arthur, Francesco Ciannella, Mingxing Du, Elin Larsen, Danny Merkx, Rachid Riad, Liming Wang, and Emmanuel Dupoux. Linguistic unit discovery from multi-modal inputs in unwritten languages: Summary of the "Speaking Rosetta" JSALT 2017 workshop. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.
- [86] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2019.
- [87] Joan Serrà, Santiago Pascual, and Carlos Segura Perales. Blow: a single-scale hyperconditioned flow for non-parallel raw-audio voice conversion. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2019.
- [88] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE, 2018.
- [89] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *Proc. International Conference on Machine Learning (ICML)*, 2018.
- [90] Yannis Stylianou, Olivier Cappé, and Eric Moulines. Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6(2):131–142, 1998.
- [91] Dídac Surís, Adrià Recasens, David Bau, David Harwath, James Glass, and Antonio Torralba. Learning words by drawing images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [92] Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. Learning words from images and speech. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2014.
- [93] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani. Voiceloop: Voice fitting and synthesis via a phonological loop. *arXiv preprint arXiv:1707.06588*, 2017.
- [94] Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura. VQVAE unsupervised unit discovery and multi-scale code2spec inverter for zerospeech challenge 2019. *arXiv preprint arXiv:1905.11449*, 2019.
- [95] Tomoki Toda, Alan W Black, and Keiichi Tokuda. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech and Language Processing*, 15(8):2222–2235, 2007.

- [96] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.
- [97] Jean-Marc Valin and Jan Skoglund. Lpcnet: Improving neural speech synthesis through linear prediction. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [98] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2017.
- [99] Christophe Veaux, Junichi Yamagishi, and Kirsten Macdonald. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit. 2017.
- [100] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [101] Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
- [102] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
- [103] Weiran Wang, Qingming Tang, and Karen Livescu. Unsupervised pre-training of bidirectional speech encoders via masked reconstruction. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.
- [104] Xin Wang, Shinji Takaki, and Junichi Yamagishi. Neural source-filter-based waveform model for statistical parametric speech synthesis. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [105] Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif Saurous. Tacotron: Towards end-to-end speech synthesis. In *Proc. Annual Conference of International Speech Communication Association (INTERSPEECH)*, 2017.
- [106] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*, 2018.
- [107] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research (JMLR)*, 2009.
- [108] Sean Welleck, Ilya Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *Proc. International Conference on Learning Representations (ICLR)*, 2020.
- [109] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. International Conference on Machine Learning (ICML)*, 2015.
- [110] Shan Yang, Yuxuan Wang, and Lei Xie. Adversarial feature learning and unsupervised clustering based speech synthesis for found data with acoustic and textual noise. *arXiv preprint arXiv:2004.13595*, 2020.
- [111] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2019.
- [112] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [113] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2014.
- [114] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

## A Visually-Grounded Speech Datasets

Table A1 displays details of the three visually-grounded speech datasets used in this paper, and distributions of utterance duration are illustrated in Figure A1. When computing duration statistics, we exclude utterances longer than 15s for SpokenCOCO and Flickr8k Audio, and 40s for Places Audio, because we found that those utterances resulted from incorrect operation of the data collection interface (e.g., workers forgot to stop recording). When computing vocabulary sizes and word statistics, text transcripts are normalized by lower-casing all the alphabets and removing characters that are neither alphabets nor digits.

For the SpokenCOCO data collection on Amazon Mechanical Turk, we displayed the text of a MSCOCO caption to a user and asked them to record themselves reading the caption out loud. For quality control, we ran a speech recognition system in the background and estimated the word-level transcription for each recording. We computed the word error rate of the ASR output against the text that the user was prompted to read, and only accepted the caption if the word error rate was under 30%. In the case that the word error rate was higher, the user was asked to re-record their speech. We paid the users \$0.015 per caption recorded, which in conjunction with the 20% overhead charged by Amazon resulted in a total collection cost of \$10,898.91.

Table A1: Statistics and properties of the three visually-grounded speech datasets used in the paper.

	SpokenCOCO	Flickr8k Audio [32]	Places Audio [36]
Num. of Utterances	605495	40000	400000
Num. of Speakers	2353	183	2683
Num. of Images	123287	8000	400000
Num. of Utterances / Image	5	5	1
Utterance Duration $\mu$	4.12s	4.33s	8.37s
Utterance Duration $\sigma$	1.31s	1.33s	4.53s
Avg. Num. of Words / Utterance	10.45	10.81	19.29
Avg. Num. of Words / Second	2.41	2.63	2.31
Total Duration	742hr	46hr	936hr
Vocabulary Size	29539	8718	41217
Type	scripted	scripted	spontaneous

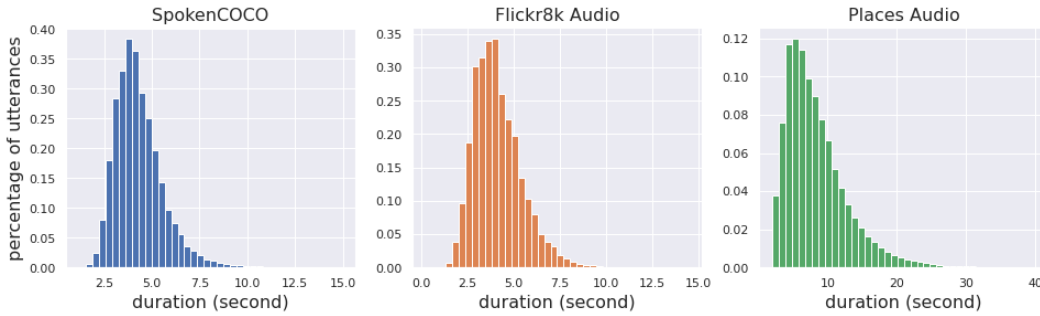


Figure A1: Utterance duration histograms for the three visually-grounded speech datasets.

## B Detailed Explanation for M-SPICE

SPICE computes an F-score between two bags of semantic propositions  $T(S)$  and  $T(c)$  parsed from a set of references  $S = \{s_i\}_i$  and a hypothesis  $c$ , where  $T(c)$  denotes a bag of propositions extracted from a scene graph parsed  $c$ , and we can compute that for multiple sentences with  $T(S) = \cup_i T(s_i)$ , and we have  $|T(S)| \geq |T(s_i)|$  for all  $i$ , as captions may capture different aspects of the same image.

To extend SPICE for scoring multiple hypotheses  $C = \{c_j\}_{j=1}^J$ , one can compute an average SPICE:  $\frac{1}{J} \sum_j F1(T(S), T(c_j))$ , or use the oracle SPICE proposed in [101]:  $\max_j F1(T(S), T(c_j))$ .



However, these metrics fail to capture the diversity among hypotheses. Let us now consider two hypothesis set,  $C^1 = \{c_1^1, c_2^1\}$  and  $C^2 = \{c_1^2, c_2^2\}$ , where

- $T(c_1^1) = T(c_2^1) = T(c_1^2) = \{(\text{girl}), (\text{table}), (\text{girl, sit-at, table})\}$
- $T(c_2^2) = \{(\text{girl}), (\text{girl, young})\}$
- $T(S) = \{(\text{girl}), (\text{table}), (\text{girl, young}), (\text{girl, sit-at, table})\}$

We would like a metric to score  $C^2$  higher than  $C^1$  *because it captures diverse and correct concepts*; however, since F1 scores are the same for  $c_1^1, c_2^1, c_1^2$ , and is lower for  $c_2^2$ , the average SPICE of  $C^1$  is higher while the oracle SPICE are the same for both  $C^1$  and  $C^2$ . Our proposed M-SPICE can be formulated as  $F1(\cup_i T(s_i), \cup_j T(c_j))$ . When hypotheses capture *diverse AND correct* propositions, the M-SPICE recall should increase as shown in Figure A2 row 2. Note that the M-SPICE score and recall will not increase if propositions are diverse but incorrect.

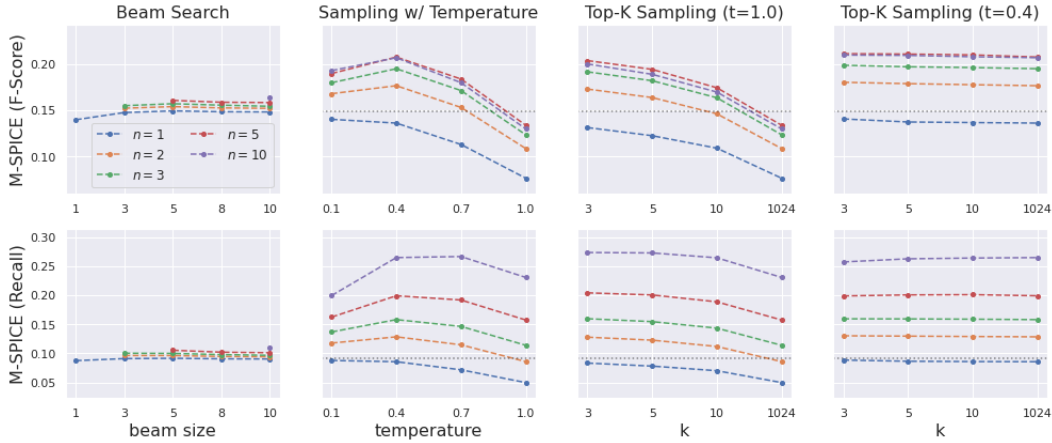


Figure A2: M-SPICE F-score (same as Figure 5) and recall on the SpokenCOCO test set with different candidate proposal methods.

## C Detailed Experimental Setups

In this section, we provide details about data preprocessing, model architecture, and training hyper-parameters for each module used in this paper. The same setups are used for all unit types unless otherwise stated.

### C.1 Image-to-Unit Model

**Data** Images are reshaped to  $256 \times 256 \times 3$  matrices and per-channel normalized with  $\mu = [0.485, 0.456, 0.406]$  and  $\sigma = [0.229, 0.224, 0.225]$ . During training, unit sequences are truncated or padded to the target length shown in Table A2. The target lengths are determined such that there are less than 10% sequences truncated while still allowing a reasonable batch size to be used. Units that occurred less than five times are excluded. Sequences are not truncated during evaluation. We follow the data splits used in [35] for Places, and [59] for Flickr8k and SpokenCOCO (commonly known as the “Karpathy split”).

Table A2: Configuration for each type of units used in the Image-to-Unit model.

	Word	Char	VQ3	VQ2	WVQ	VQ3 \ RLE
Target Length	18	70	100	200	110	160
Sequence Truncated (%)	1.12	1.74	6.90	9.37	7.80	6.35
Batch Size (SAT)	80	60	40	40	40	40
Batch Size (SAT-FT)	32	32	20	-	-	-

**Model** We adopt an open-source re-implementation<sup>3</sup> of Show, Attend, and Tell [109] (SAT) with soft attention, which replaces the CNN encoder in [109] with a ResNet-101 [41] pre-trained on ImageNet [19] for image classification. The last two layers of the ResNet are removed (a pooling layer and a fully-connected layer) such that the encoder produces a  $14 \times 14 \times 2048$  feature map for each image.

**Training** Two model variants are considered in this paper: SAT and SAT-FT, which differ in how each part is initialized and which parts are updated during training. The SAT model initializes the encoder parameters with a pre-trained image classification model and freezes the encoder parameters during training. On the other hand, the SAT-FT model (fine-tuned SAT model) initializes the entire model with a pre-trained SAT model, and update all parameters during training. Adam [61] with a learning rate of  $10^{-4}$  is used for optimizing both models. The training objective is maximum likelihood combined with a doubly stochastic attention regularization introduced in [109] with a weight of 1. Dropout is applied to the input of decoder softmax layer with a probability of 0.5 during training. Gradients are clipped at 5 for each dimension. The batch size for each unit is shown in Table A2, which are chosen based on the target length and the GPU memory constraints. All SAT models are trained for at most 30 epochs, and SAT-FT models are trained for at most 20 epochs. Models are selected based on the unit BLEU-4 score on the validation set.

The time complexity of forward computation is the same for the encoder for all units, while for the decoder it is proportional to the unit sequence length due to the auto-regressive nature. Using two NVIDIA TITAN X Pascal GPUs with data parallel training, each epoch takes about 2.8 hours for VQ3 units and 5.3 hours for VQ2 units.

## C.2 Unit-to-Speech Model

**Data** Run-length encoded unit sequences are used as input for all systems (i.e., VQ3 and VQ3 \ RLE systems share the same unit-to-speech model). The native sample rates of audio files in LJSpeech [55] and VCTK [99] are 22050Hz and 48kHz, respectively. For consistency and compatibility with the spectrogram-to-waveform model, we down-sample those in VCTK to 22050Hz. Following Tacotron [105] and Tacotron2 [88], we compute a 80 dimensional Mel spectrogram for each audio file with a 256-sample frame hop, a 1024-sample frame size, and a Hann window function. At a sample rate of 22050Hz, it corresponds to about 11.6ms frame hop and 46.4ms frame size. Utterances longer than 8 seconds are discarded during training to accommodate for the GPU memory constraints. We follow the data splits provided at <https://github.com/NVIDIA/tacotron2> for LJSpeech. For the multi-speaker VCTK dataset, to ensure the same speaker distribution between train and valid splits, we randomly sample 2.5% of the utterances from each speaker for validation, which results in a set of 1087 utterances.

**Model** We use a cascade of two systems to convert a unit sequence into a waveform. The first part synthesizes a Mel spectrogram from a sequence of units, which is referred to as the Unit-to-Speech model in this paper and determines most of the properties of interest in speech (e.g., linguistic content, speaker, prosody). The second part is a vocoder that converts a Mel spectrogram into a waveform, which mostly affects the fidelity of the synthesized speech rather than the aforementioned properties. A vocoder can be either a learned module like WaveNet [75] and WaveGlow [81], or a parameter-less signal processing block such as using the Griffin-Lim algorithm [28].

We use an re-implementation<sup>4</sup> of Tacotron2 [88] for Unit-to-Speech models, which is a sequence-to-sequence encoder-decoder model with location-sensitive attention [13]. For single-speaker models trained on the LJSpeech dataset, the exact same hyperparameters and model architecture are used as [88]. For multi-speaker models trained on the VCTK dataset, we create an additional speaker embedding table of 256 dimensions for all speakers and control the speaker identity through these speaker embeddings. Speaker embeddings are injected at two places in the decoder: first in concatenation with the original input to the decoder LSTM, and second in concatenation with the output of the decoder LSTM, right before predicting the stop token and the spectra of a frame.

<sup>3</sup><https://github.com/sgrvinod/a-PyTorch-Tutorial-to-Image-Captioning>

<sup>4</sup><https://github.com/NVIDIA/tacotron2>

A pre-trained<sup>5</sup> WaveGlow [81] vocoder is used for all Unit-to-Speech models, which demonstrates the universality of vocoder models and how little acoustic properties of interest are affected by them. Although it is possible to achieve a even higher fidelity score through training or fine-tuning the WaveGlow model on the re-synthesized spectrograms, we did not attempt to experiment with that, since the focus of this paper is to demonstrate the capability of generating fluent spoken captions and controlling properties like speaker identity independently.

**Training** A batch size of 64 are used for all systems. Adam [61] with an initial learning rate of  $10^{-3}$  is used to minimize the mean square error from spectrogram prediction and the binary cross entropy from stop token prediction combined. L2 regularization for the parameters with a weight of  $10^{-6}$  is applied, and the L2 norm of the gradients are clipped at 1. Models are trained for 500 epochs on LJSpeech and 250 epochs on VCTK, and selected based on the validation loss.

The time complexity of forward computation at the encoder is proportional to the unit-sequence length because of the bi-directional LSTM layer in the encoder. On the other hand, the number of the decoding steps is proportional to the duration of the speech at the decoder, which is independent of the choice of input representation; however, at each decoding step, the number of encoder outputs the attention module attends to is proportional to length of the unit sequence. Empirically, each training epoch on LJSpeech takes about 12 minutes using two NVIDIA Titan X Pascal GPUs for both VQ2 and VQ3 models despite that VQ2 sequences are in average twice as long as VQ3 sequences, which shows that time complexity are dominated by other computations.

### C.3 Speech-to-Unit Model

We obtain the ResDAVENet-VQ “{2} → {2, 3}” model and the WaveNet-VQ (PA) model reported in [35] from the authors. Both models learn discrete representations for speech and are used to transcribe speech into a sequence of units in this paper. We use these models to extract unit sequences for all datasets without fine-tuning, which examines the robustness of these Speech-to-Unit models when applied to datasets of different domains. Table A3 compares the three types of units used in this paper (VQ3, VQ2, WVQ) extracted from these two models. For self-containedness, the ABX error rate for each unit reported in [35] are also included. The ABX test evaluates the phone discriminability of the learned units on the ZeroSpeech 2019 English test set [24]. Note that (1) VQ3 and WVQ have the same unit rate before run-length encoding, (2) VQ2 achieves the lowest ABX error rate, (3) and all units have a lower ABX error rate before run-length encoding.

Table A3: Properties of the three types of units and the two speech-to-unit models.

	VQ3	VQ2	WVQ
Source Model	ResDAVENet-VQ	ResDAVENet-VQ	WaveNet-VQ
Training Data	Places Audio+Image	Places Audio+Image	Places Audio
Training Objective	Contrastive Loss	Contrastive Loss	Reconstruction Loss
RLE ABX Error Rate	15.68%	13.06%	25.23%
Pre-RLE ABX Error Rate	14.52%	12.51%	24.87%
Pre-RLE Unit Rate	40ms	20ms	40ms

## D Image-to-Unit Samples

Table A4 and A5 display unit captions for the same image generated from Image-to-Unit models trained on different learned units. Captions in Table A4 are decoded with beam search (beam size=5), and those in Table A5 are sampled from the model distribution with top-k sampling ( $k = 5$ ).

Table A4 shows that Image-to-Unit models trained on WVQ and VQ3 units without run-length encoding (VQ3 \ RLE) fail to produce reasonable captions using beam search for *all* images. In fact, generated captions are almost always the same among different images for these two models. The WVQ model generates captions looping the same bi-gram until exceeding the maximum length, while the VQ3 \ RLE repeats the same unit. On the other hand, the model trained on VQ2 units can sometimes produce reasonable captions, but it exhibit the same behavior as the WVQ model when it fails as shown here.

<sup>5</sup><https://github.com/NVIDIA/waveglow>

On the contrary, Table A5 shows that all four models are capable of generating non-trivial captions without looping via sampling.<sup>6</sup> The observation here is consistent with the evaluation results on the transcribed spoken captions presented in Table A7 and A8 and Figure A3.

Table A4: Exemplar beam search decoding results from SAT Image-to-Unit models.

Symbol	Captioned Generated with Beam Search (beam size=5)
VQ3	263 32 208 5 336 100 717 803 256 803 815 144 120 144 654 936 48 417 272 417 362 766 825 284 614 156 341 135 769 5 208 32 208 5 336 815 144 815 494 181 467 417 870 395 683 141 250 543 820 587 181 913 1013 467 5 208 32 208 5 467 360 606 360 801 1009 398 847 89 100 869 254 1003 442 42 791 417 272 141 766 362 614 156 341 135 769 5 208 32
VQ2	71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791 71 791...
WVQ	181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232 181 232...
VQ3 \ RLE	263 32...

Table A5: Exemplar sampling results with  $(t, k) = (1.0, 5)$  from SAT Image-to-Unit models.

Symbol	Captioned Generated with top-k sampling ( $k = 5$ )
VQ3	263 208 467 717 288 426 986 72 44 341 151 801 1022 27 320 426 288 66 570 683 351 313 910 820 543 820 230 100 852 329 852 288 502 706 427 110 451 297 938 457 426 100 852 329 852 791 993 522 993 374 502 288 936 48 263 208 32
VQ2	71 791 71 791 71 791 71 191 175 51 139 359 173 599 307 419 133 621 85 165 315 883 175 191 71 791 71 48 511 765 983 873 314 409 333 267 409 734 229 787 184 937 886 254 934 666 973 19 947 227 805 967 883 175 48 695 511 655 806 491 647 507 343 867 819 655 699 491 136 221 513 996 675 581 467 652 488 186 3 183 311 613 371 463 314 21 238 910 238 657 230 82 270 868 643 78 391 940 922 49 771 986 147 947 19 957 862 957 95 7 819 695 1011 159 831 589 966 827 753 891 162 253 269 219 13 501 977 302 241 157 691 723 695 175 191 71 791 71 791 71 791 71 48 1007
WVQ	181 232 181 232 181 232 181 232 181 232 181 225 124 232 181 232 225 232 181 225 124 225 232 181 252 169 211 147 89 67 156 155 189 110 53 246 225 89 52 21 5 216 155 225 25 47 41 223 225 181 166 57 185 82 25 225 124 149 214 93 28 195 65 1 23 109 246 223 141 47 41 223 181 232 82 231 188 169 147 89 225 181 225 124 181 124 5 216 53 246 181 225 137 52 5 159 225 181 225 46 155 246 232 181 232 225 232 181 225 52 30 5 216 166 225 124 225 181 225 5 4 46 225 181 25 137 52 159 155 225 181 225 108 155 246 225 108 155 225 232 181 25 89 221 70 197 232 181 225 214 28 214 225 181 232 244 220
VQ3 \ RLE	263 32 870 870 870 968 910 250 543 820 587 909 909 181 717 48 936 48 224 176 284 538 133 807 715 39 27 27 476 5 5 476 570 395 395 683 313 141 250 250 587 587 494 909 922 181 100 100 827 119 66 272 417 766 766 766 614 614 156 341 135 135 181 913 913 1013 5 208 208 208 208 208 208 5 5 5 476 320 96 96 651 538 133 766 766 825 740 913 1013 467 5 208 208 32 32 32 32 208 208 5 5 336 501 254 254 254 1003 442 852 362 825 740 639 639 587 543 543 975 320 320 284 284 228 844 844 622 622 846 654 654 846 336 263 208 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32

<sup>6</sup>The model might still generate trivial captions when sampling with a temperature close to zero or setting a very small  $k$  with top-k sampling, in which case sampling is similar to greedy decoding.

## E Caption Evaluation on Unit Sequences

The spoken caption evaluations presented in Section 3.2 utilized an automatic speech recognizer to transcribe the generated captions into words so that they could be compared against reference text captions. In the case that a speech recognizer is not available, we can perform evaluation directly on the speech unit representations by using the unit model to transcribe the reference spoken captions. Table A6 displays BLEU-4, METEOR, ROUGE, and CIDEr scores evaluated directly on the various speech unit representations; we cannot compute SPICE scores directly on the speech units because SPICE requires a dependency tree for each caption. It is important to note that for a given evaluation metric, the scores across the models are not directly comparable because their unit spaces are different. We do note that the relative ranking among VQ3, VQ2, and Wavenet-VQ is consistent across BLEU-4, METEOR, and ROUGE, however, VQ3 \ RLE achieves abnormally high scores on these metrics despite producing trivial captions for all images as shown in Table A4. This is because unit “32” has learned to represent non-speech frames such as silence, which frequently occurs at both the beginning and end of utterances. Without RLE, consecutive strings of “32” units are extremely common in both the candidate and reference captions, which inflates the scores of this model. The exception here is the CIDEr metric, which incorporates TF-IDF weighting that tends to de-emphasize these kinds of uninformative patterns. The fact that the CIDEr score is 0 for both the VQ3 \ RLE and Wavenet-VQ models indicates that in general the captions produced by these models are uninformative. We posit that word-level evaluation is always preferable for spoken caption generation, but in the case that this is not possible the CIDEr metric may be the best option.

Table A6: Unit-based caption evaluation on MSCOCO test set. The beam size  $\in \{3, 5, 10\}$  was chosen for each model to maximize the CIDEr score. Note that the scores between different units are not directly comparable, because they are computed based different types of units.

symbol	Greedy / Beam-Search (SAT Model)			
	Unit BLEU-4	Unit METEOR	Unit ROUGE	Unit CIDEr
VQ3	0.176 / 0.274	0.178 / 0.196	0.280 / 0.328	0.121 / 0.215
VQ2	0.172 / 0.141	0.132 / 0.108	0.178 / 0.157	0.027 / 0.020
WVQ	0.019 / 0.020	0.048 / 0.048	0.081 / 0.081	0.000 / 0.000
VQ3 \ RLE	0.163 / 0.163	0.168 / 0.168	0.218 / 0.218	0.000 / 0.000

## F Full Results of Caption Evaluation on Word Sequences

Table A7 and A8 and Figure A3 present the complete caption evaluation results on transcribed word sequences for all 5 metrics, supplementing Figure 3 in the main paper that only presents the SPICE results. Note that the transcripts for WVQ and VQ3 \ RLE beam search captions are not reliable; for the reasons discussed in the previous section the generated spoken captions contain only silence, and the ASR model used for transcription did not see utterances comprised of pure silence during training. We see that ranking between symbols are generally consistent among all those metrics, except the ranking between WVQ and VQ3 \ RLE when sampling with a temperature of 0.4. This is a relatively low-score regime when both model are transiting from generating trivial caption ( $t = 0.1$ ) to non-trivial captions ( $t = 0.7$ ).

Table A7: Word-based caption evaluation on MSCOCO test set. An ASR model is used to transcribe the spoken captions into text for evaluation. The beam size  $\in \{3, 5, 10\}$  was chosen for each model to maximize the SPICE score.

symbol	Greedy / Beam-Search (SAT Model)				
	BLEU-4	METEOR	ROUGE	CIDEr	SPICE
word	0.287 / 0.315	0.247 / 0.253	0.524 / 0.533	0.939 / 0.984	0.180 / 0.185
char	0.238 / 0.289	0.230 / 0.239	0.495 / 0.512	0.783 / 0.879	0.164 / 0.172
VQ3	0.133 / 0.186	0.162 / 0.186	0.413 / 0.446	0.435 / 0.584	0.111 / 0.127
VQ2	0.068 / 0.073	0.138 / 0.126	0.343 / 0.345	0.262 / 0.224	0.084 / 0.065
WVQ	0.010 / 0.009	0.069 / 0.069	0.286 / 0.285	0.009 / 0.009	0.011 / 0.011
VQ3 \ RLE	0.000 / 0.000	0.002 / 0.002	0.001 / 0.001	0.000 / 0.000	0.001 / 0.001

Table A8: Sampling-based evaluations with SAT models trained on different units.

Metric	symbol	Sampling with Temperature				Top-K Sampling ( $t = 1.0$ )			Top-K Sampling ( $t = 0.7$ )		
		$t = 1.0$	$t = 0.7$	$t = 0.4$	$t = 0.1$	$k = 10$	$k = 5$	$k = 3$	$k = 10$	$k = 5$	$k = 3$
BLEU-4	VQ3	0.052	0.097	0.132	0.137	0.084	0.108	0.120	0.109	0.119	0.124
	VQ2	0.039	0.058	0.068	0.066	0.059	0.068	0.069	0.064	0.070	0.071
	WVQ	0.033	0.047	0.025	0.012	0.056	0.050	0.037	0.052	0.042	0.025
	VQ3 \ RLE	0.049	0.075	0.035	0.000	0.070	0.087	0.092	0.082	0.094	0.093
METEOR	VQ3	0.124	0.151	0.168	0.165	0.147	0.160	0.166	0.159	0.165	0.168
	VQ2	0.115	0.134	0.146	0.140	0.134	0.142	0.147	0.140	0.144	0.147
	WVQ	0.096	0.106	0.078	0.069	0.112	0.104	0.088	0.105	0.094	0.080
	VQ3 \ RLE	0.119	0.135	0.055	0.002	0.136	0.146	0.148	0.141	0.144	0.141
ROUGE-L	VQ3	0.303	0.358	0.403	0.416	0.346	0.371	0.386	0.373	0.386	0.397
	VQ2	0.293	0.330	0.351	0.345	0.325	0.345	0.351	0.340	0.348	0.355
	WVQ	0.270	0.297	0.287	0.287	0.312	0.309	0.292	0.309	0.295	0.276
	VQ3 \ RLE	0.295	0.330	0.152	0.001	0.328	0.349	0.355	0.340	0.348	0.350
CIDEr	VQ3	0.195	0.345	0.461	0.451	0.312	0.383	0.424	0.395	0.431	0.444
	VQ2	0.143	0.231	0.272	0.267	0.220	0.260	0.277	0.251	0.270	0.278
	WVQ	0.095	0.150	0.044	0.009	0.180	0.145	0.082	0.154	0.116	0.055
	VQ3 \ RLE	0.182	0.277	0.130	0.000	0.260	0.316	0.340	0.304	0.328	0.332
SPICE	VQ3	0.063	0.093	0.111	0.114	0.086	0.100	0.108	0.100	0.106	0.109
	VQ2	0.052	0.074	0.086	0.087	0.073	0.082	0.085	0.079	0.084	0.087
	WVQ	0.035	0.046	0.019	0.011	0.051	0.042	0.026	0.043	0.034	0.020
	VQ3 \ RLE	0.060	0.078	0.034	0.001	0.077	0.087	0.091	0.083	0.088	0.086

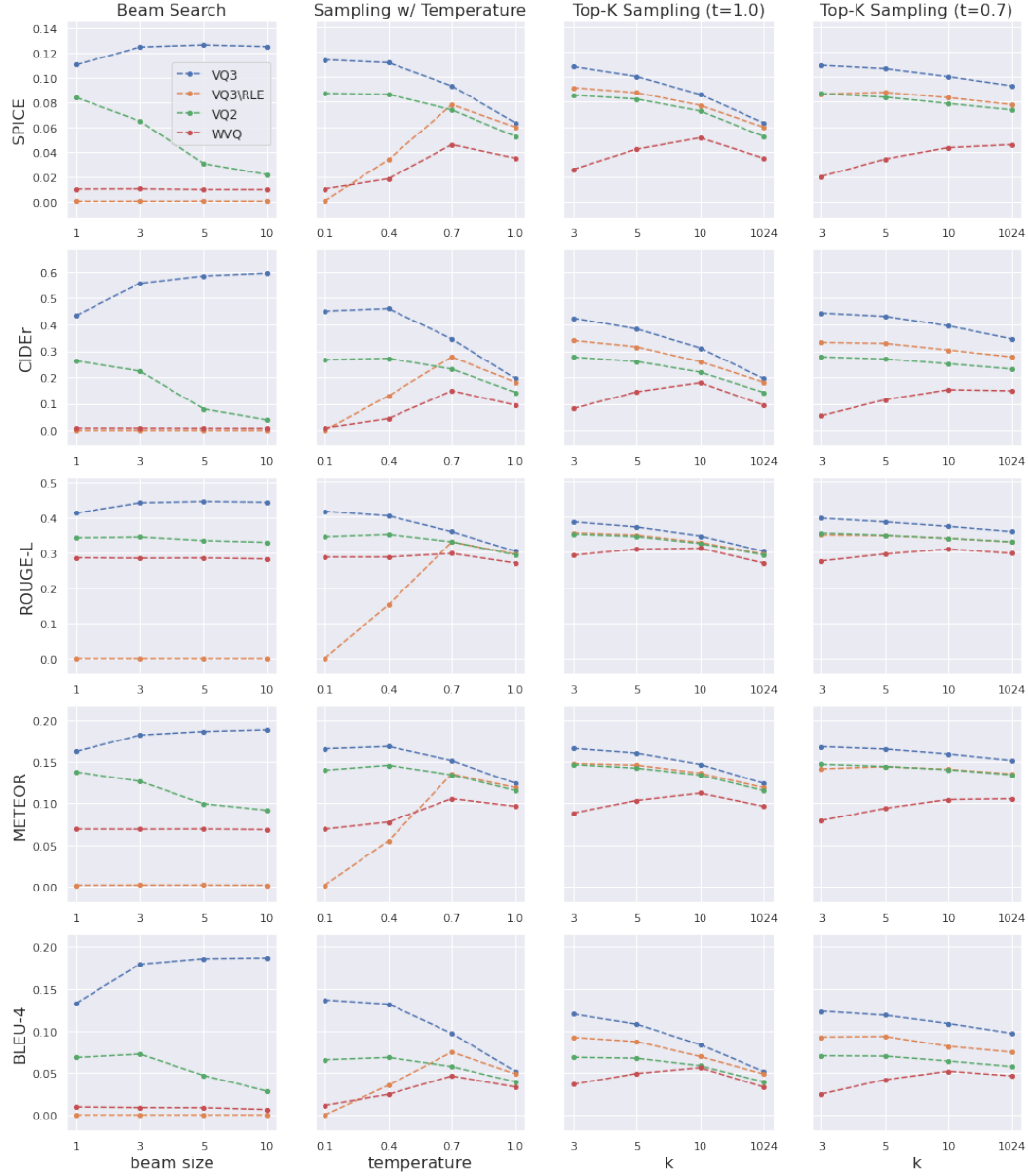


Figure A3: Word-based caption evaluation results of all five metrics on SAT models trained on different units, using both beam search decoding and sampling for unit sequence generation.



## G Comparison of Average SPICE and M-SPICE

Figures A4 and A5 display the M-SPICE scores and SPICE score distributions of different sampling methods for the SAT and SAT-FT model trained on VQ3 units, respectively. The exact numbers are shown in Tables A9, A11, A10, and A12. When performing sampling-based evaluation, there is bound to be some stochasticity in the results. However, the SPICE score distributions (box plots over 10 trials) shown in the bottom row of Figures A4 and A5 are very narrow, which we attribute to the fact that the COCO test set is large enough attenuate this stochasticity. The narrowness of the box plots also suggests that taking the average SPICE score over multiple sampling runs does not reflect the diversity of the captions the way that M-SPICE does.

Table A9: M-SPICE F1-scores of the VQ3 SAT model with beam search decoding.

$n =$	Beam Size				
	1	3	5	8	10
1	0.111	0.125	0.127	0.126	0.125
2	-	0.127	0.130	0.129	0.128
3	-	0.129	0.131	0.131	0.130
5	-	-	0.134	0.133	0.132
10	-	-	-	-	0.135

Table A10: M-SPICE F1-scores of the VQ3 SAT-FT model with beam search decoding.

$n =$	Beam Size				
	1	3	5	8	10
1	0.140	0.147	0.149	0.148	0.148
2	-	0.152	0.154	0.152	0.152
3	-	0.155	0.157	0.155	0.154
5	-	-	0.161	0.159	0.158
10	-	-	-	-	0.164

Table A11: M-SPICE F1-scores of the VQ3 SAT model with sampling.  $t$  denotes the temperature, and  $k$  denotes the number of top units considered at each decoding step for top-K sampling.

$n =$	Sampling with Temperature				Top-K Sampling ( $t = 1.0$ )			Top-K Sampling ( $t = 0.4$ )		
	$t = 1.0$	$t = 0.7$	$t = 0.4$	$t = 0.1$	$k = 10$	$k = 5$	$k = 3$	$k = 10$	$k = 5$	$k = 3$
1	0.063	0.093	0.111	0.114	0.086	0.100	0.108	0.100	0.106	0.109
2	0.092	0.128	0.147	0.137	0.120	0.137	0.145	0.138	0.143	0.146
3	0.106	0.145	0.163	0.147	0.137	0.153	0.161	0.153	0.160	0.163
5	0.117	0.156	0.175	0.154	0.148	0.165	0.173	0.164	0.173	0.174
10	0.115	0.153	0.173	0.155	0.145	0.160	0.169	0.162	0.169	0.171

Table A12: M-SPICE F1-scores of the VQ3 SAT-FT model with sampling.

$n =$	Sampling with Temperature				Top-K Sampling ( $t = 1.0$ )			Top-K Sampling ( $t = 0.4$ )		
	$t = 1.0$	$t = 0.7$	$t = 0.4$	$t = 0.1$	$k = 10$	$k = 5$	$k = 3$	$k = 10$	$k = 5$	$k = 3$
1	0.076	0.113	0.136	0.140	0.109	0.122	0.131	0.137	0.137	0.141
2	0.108	0.153	0.177	0.168	0.146	0.164	0.173	0.178	0.179	0.180
3	0.123	0.171	0.195	0.180	0.163	0.182	0.192	0.196	0.197	0.199
5	0.134	0.184	0.208	0.190	0.174	0.194	0.204	0.210	0.211	0.211
10	0.130	0.180	0.207	0.193	0.170	0.189	0.200	0.208	0.209	0.210

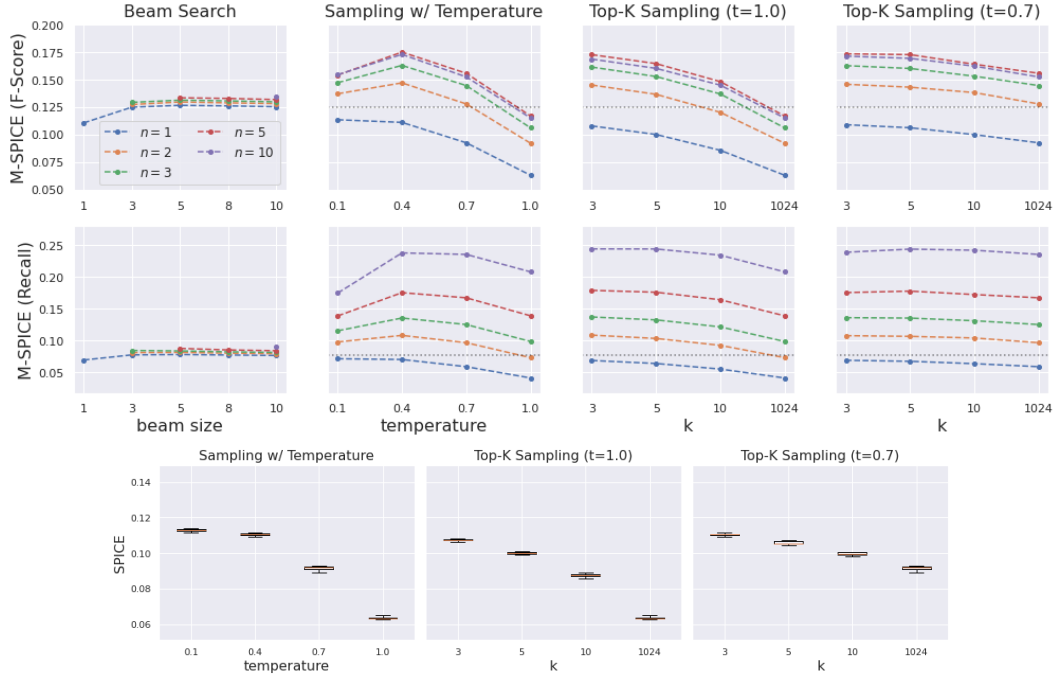


Figure A4: M-SPICE F1-scores and recalls (top) and SPICE distributions (bottom) of the SAT model on the MSCOCO test set with different caption generation methods. Box-and-whisker plots the SPICE scores over 10 runs are shown, where a box extends from the first quartile to the third quartile.

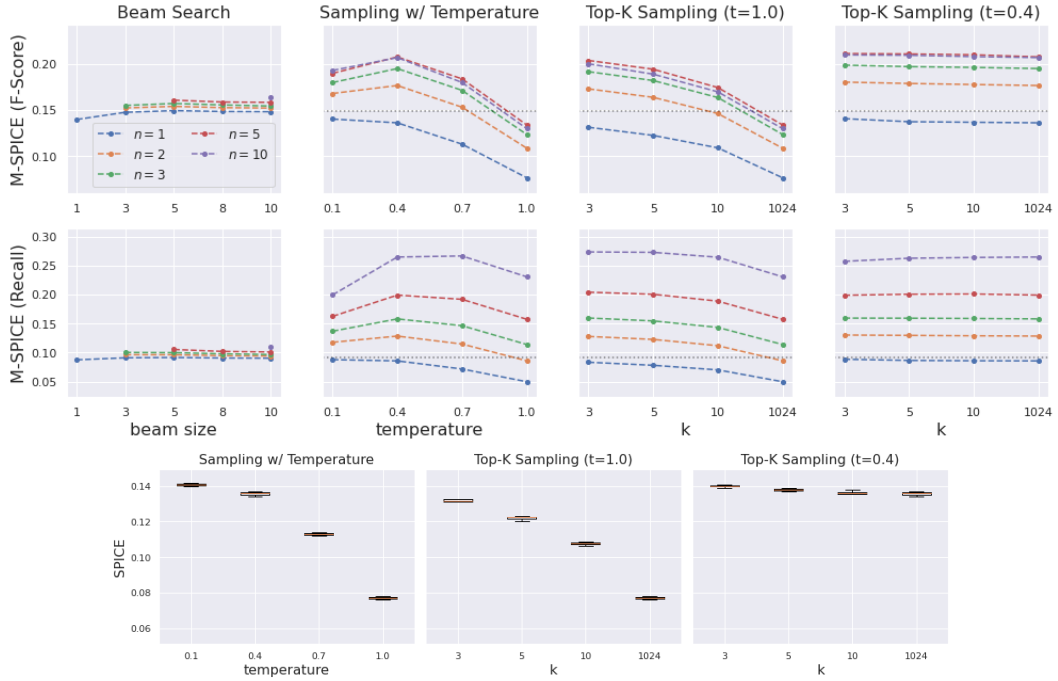


Figure A5: M-SPICE F1-scores and recalls (top) and SPICE distributions (bottom) of the SAT-FT model on the MSCOCO test set with different caption generation methods.

## H Full Results of Learned Vocabulary Size

In Table A13, we display the numerical results depicted graphically in Figure 4.

Table A13: The vocabulary size of the VQ3 SAT-FT model as estimated by various decoding approaches. The numbers in this table display the specific values of the curves depicted in Figure 4.

$n$	Beam Search beam size=?					Sampling ( $t$ : temperature; $k$ : top-k)									
	1	3	5	8	10	$(t, k) = (?, All)$				$(t, k) = (1.0, ?)$			$(t, k) = (0.7, ?)$		
1	551	479	447	421	411	1447	978	689	561	1058	908	770	694	663	670
2	-	572	523	502	474	2100	1367	917	696	1522	1289	1025	907	867	851
3	-	693	620	585	562	2550	1644	1075	803	1855	1515	1222	1069	1003	973
5	-	-	681	625	617	3239	2111	1305	938	2367	1861	1511	1266	1209	1155
10	-	-	-	-	700	4311	2876	1664	1155	3176	2512	1954	1618	1552	1437

## I Disentangled Voice Control for Image-to-Speech Synthesis

We examine to what extent the VQ3 units are portable across different speakers by training a U2S model on the VCTK dataset that additionally takes a speaker ID as input. The resulting model is able to generate speech with the voice of any VCTK speaker. We evaluate the captions produced by this system on SpokenCOCO for 5 speakers in Table A14. In order to compute these scores we transcribe the captions generated by each model into text using the ASR system we describe in Section 4, which was solely trained on re-synthesized SpokenCOCO captions using the LJSpeech U2S model. The scores in Table A14 indicate not only that the I2U model can be easily integrated with U2S models representing a diverse set of speakers, but also that the LJSpeech ASR system works very well on the speech synthesized from the VCTK models. In Figure 1, we show example captions generated by conditioning on the same unit sequence, but different speaker identities.

Table A14: Demonstration of disentangled voice control via synthesizing the same units with different unit-to-speech models conditioned on different speaker IDs. Units are generated with beam search using the SAT-FT model for the MSCOCO test set.

Train Data	Speaker ID	Gender	Region	BLEU-4	METEOR	ROUGE	CIDER	SPICE
LJSpeech	-	F	-	0.233	0.212	0.478	0.732	0.149
VCTK	p247	M	Scottish	0.234	0.211	0.480	0.730	0.148
	p231	F	English	0.233	0.210	0.478	0.724	0.146
	p294	F	American	0.236	0.212	0.482	0.732	0.148
	p345	M	American	0.234	0.209	0.477	0.717	0.144
	p307	F	Canadian	0.234	0.211	0.479	0.729	0.148

## J More Image-to-Speech Samples

In Tables A15 and A16, we show many more examples of spoken captions generated by the VQ3 model. In Table A15, all three captions in each row were generated from the same unit sequence corresponding to the top hypothesis from beam search decoding. Each column represents the caption waveform generated by a different U2S model reflecting a different speaker. Although the spectrograms are visibly very different (reflecting the differing speaker characteristics across the U2S models), the word sequence estimated by the ASR model is generally the same. We do notice some substitution errors (highlighted in red), most commonly between “a” and “the”.

Table A16 shows captions generated by the same VQ3 model and for the same set of images depicted in Table A15, but instead of varying the U2S model we show captions generated via sampling rather than beam search. Here, we note that the sampled captions exhibit diversity both their content and linguistic style. We observe that the captioning model has learned to produce captions that correctly use quantifiers and conjugate verbs (“a couple of cows walking” vs. “a cow is standing”). The model also disentangles object identity from attributes such as color “red fire hydrant” vs. “yellow fire hydrant” vs. “green fire hydrant”).

Table A15: Samples. More at [https://wnhsu.github.io/image-to-speech-demo/3\\_vq3\\_voice\\_control\\_sat-ft\\_model](https://wnhsu.github.io/image-to-speech-demo/3_vq3_voice_control_sat-ft_model)

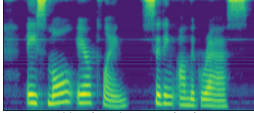
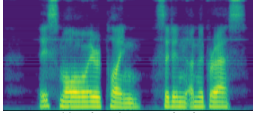
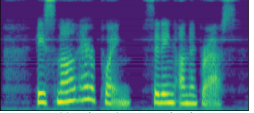

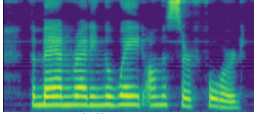
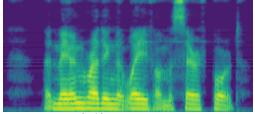
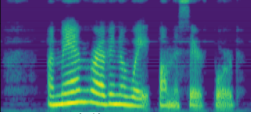

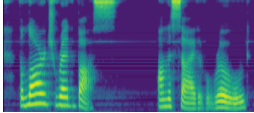
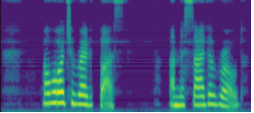

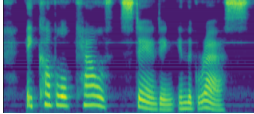
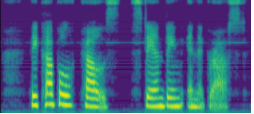

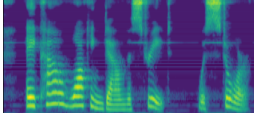
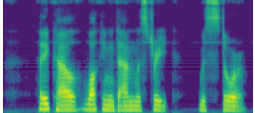
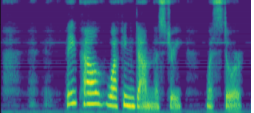

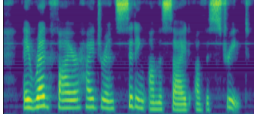
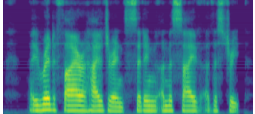
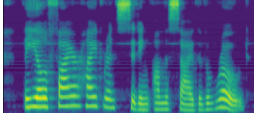
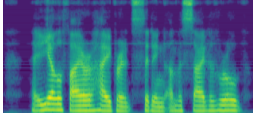
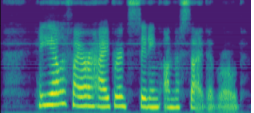
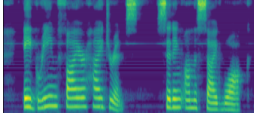
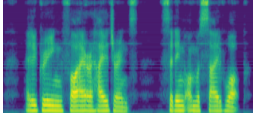
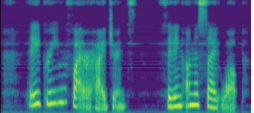
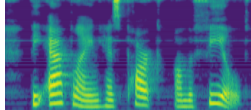
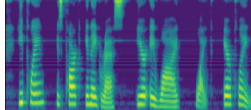
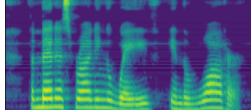
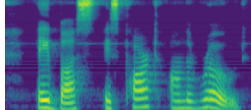

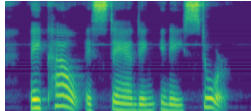

Image	Generated Spoken Captions / Transcripts (SAT-FT, VQ3, Beam Search)		
	LJSpeech	VCTK (p247)	VCTK (p307)
	 a small airplane sitting on the grass	 a small airplane sitting on the grass	 a small airplane sitting on the grass
	 a man riding a wave on a surfboard	 a man riding a wave on a surfboard	 a man riding a wave on a surfboard
	 a large red bus on the side of <b>the</b> road	 a large red bus on the side of <b>the</b> road	 a large red bus on the side of <b>a</b> road
	 a couple of cows standing in the grass	 a couple of cows standing in the grass	 a couple of cows standing in the grass
	 a cow walking down the street <b>with</b> a store	 a cow walking down the street <b>in</b> a store	 a cow walking down the street <b>next to</b> a store
	 a red fire hydrant sitting on the side of a street	 a red fire hydrant sitting on the side of a street	 a red fire hydrant sitting on the side of a street
	 a yellow fire hydrant sitting on the side of <b>a</b> road	 a yellow fire hydrant sitting on the side of <b>the</b> road	 a yellow fire hydrant sitting on the side of <b>a</b> road
	 a green fire hydrant sitting on <b>a</b> sidewalk	 a green fire hydrant sitting on <b>the</b> sidewalk	 a green fire hydrant sitting on <b>a</b> sidewalk

Table A16: Samples. More at [https://wnhsu.github.io/image-to-speech-demo/2\\_vq3\\_sample\\_diversity\\_sat-ft\\_model](https://wnhsu.github.io/image-to-speech-demo/2_vq3_sample_diversity_sat-ft_model)

Image	Generated Spoken Captions / Transcripts (SAT-FT, VQ3, Sampling $(t, k) = (0.4, 3)$ )		
trial 1	trial 2	trial 3	
	 the airplane is parked on the field	 a plane is parked in the grass near a white and white airplane	 a small airplane that is standing in a field
	 a surfer riding a wave in the water	 the man is riding the wave in the water	 a surfer is riding a wave on a wave
	 the bus parked on the side of the road	 a large red bus is stopped in the road	 a bus is parked on the road
	 a couple of cows walking in a field	 a couple of cows in a grassy field	 a couple of cows walking in a grassy field
	 a cow is standing in a store	 a brown cow walking down the side of a street	 a brown and white cow standing in a line
	 a red fire hydrant is sitting on the side of the street	 a red fire hydrant sitting on a sidewalk in a concrete	 a red fire hydrant sitting on the side of a road
	 a yellow fire hydrant in the middle of the side of a road	 a yellow fire hydrant is sitting in the park	 a yellow fire hydrant in a line on the side of a street
	 a fire hydrant on a sidewalk in the middle	 a green fire hydrant on the side of the road	 a fire hydrant with a curb on the side of the street