Multichannel Generative Language Model:
Learning All Possible Factorizations Within and Across Channels

Harris Chan[*]
Vector Institute
University of Toronto
hchan@cs.toronto.edu

Jamie Kiros
Google Research, Brain Team
kiros@google.com

William Chan
Google Research, Brain Team
williamchan@google.com

Abstract

A channel corresponds to a viewpoint or transformation of an underlying meaning. A pair of parallel sentences in English and French express the same underlying meaning, but through two separate channels corresponding to their languages. In this work, we present the Multichannel Generative Language Model (MGLM). MGLM is a generative joint distribution model over channels. MGLM marginalizes over all possible factorizations within and across all channels. MGLM endows flexible inference, including unconditional generation, conditional generation (where 1 channel is observed and other channels are generated), and partially observed generation (where incomplete observations are spread across all the channels). We experiment with the Multi30K dataset containing English, French, Czech, and German. We demonstrate experiments with unconditional, conditional, and partially conditional generation. We provide qualitative samples sampled unconditionally from the generative joint distribution. We also quantitatively analyze the quality-diversity trade-offs and find MGLM outperforms traditional bilingual discriminative models.

1 Introduction

A natural way to consider two parallel sentences in different languages is that each language expresses the same underlying meaning from a different viewpoint. Each language can be thought of as a transformation that maps an underlying concept into a view that we collectively agree is determined as 'English' or 'French'. Similarly, an image of a cat and the word 'cat' are expressing two views of the same underlying concept. In this case, the image corresponds to a high bandwidth channel and the word 'cat' to a low bandwidth channel. This way of conceptualizing parallel viewpoints naturally

[*]Work done during an internship at Google Brain.

leads to the formulation of a fully generative model over each instance, where the transformation corresponds to a particular generation of the underlying view. We define each of these views as a channel. As a concrete example, given a parallel corpus of English and French sentences, English and French become two channels, and the corresponding generative model becomes $p(\text{English}, \text{French})$. One key advantage of this formulation is that a single model can be trained to capture the full expressivity of the underlying concept, allowing us to compute conditionals and marginals along with the joint. In parallel sentences, the conditionals correspond to translations from one channel to another while the marginals correspond to standard monolingual language models.

In this work, we present a general framework for modeling the joint distribution $p(\mathbf{x}_1, ..., \mathbf{x}_k)$ over k channels by marginalizing over all possible factorizations across the channels and within each channel. This formulation allows our framework to perform: 1) unconditional generation, 2) fully conditional generation (source channels are fully observed and fixed), and 3) *partial* conditional generation (source channels contain incomplete sequences).

The key contributions in this work are:

1. We present MGLM, a multichannel generative modeling framework. MGLM models the joint distribution $p(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ over k channels by marginalizing over all possible factorization across and within sequences.

2. Since MGLM is trained over all possible factorizations, MGLM can perform both conditional generation (e.g., machine translation with fully observed source channel), and *partially observed* conditional generation across different channels (e.g., seeding each channel

4208

| "The green" | "grenouille" | "Der ist" | "" | → MGLM → | "The **frog is** green" | "La grenouille **est verte**" | "Der **Frosch** ist **grün**" | "**Žába je zelená**" |

| English | French | German | Czech | | English | French | German | Czech |

Figure 1: **Multichannel Generative Language Models** (MGLM) marginalize over all possible factorizations of the joint distribution within and across all channels (e.g., languages). MGLM is trained to predict the tokens to be inserted (in bold), given partially observed inputs. At inference, MGLM can take *full*, *partial*, or *empty* sequence from each channel and generate the full sequence for each channel.

3. In the case of conditional generation over multiple target languages, we show that we are competitive in BLEU and have significant advantages in inference time and model memory savings.

4. We analyze the Quality-Diversity tradeoff from sampling MGLM and prior work.

We highlight that while we focus on languages as a specific instantiation of a channel, our framework can generalize to any arbitrary specification, such as other types of tasks (e.g., question-answering) or other modalities (e.g., image captioning).

2 Multichannel Generative Language Model

In multichannel generative language modeling, our goal is to learn a generative model given a dataset consisting of a set of sequences $\{\mathbf{x}_1^{(i)}, \ldots, \mathbf{x}_k^{(i)}\}_{i=1}^{M}$ from up to k channels, where $\mathbf{x}_j^{(i)} = [x_{j,1}^{(i)}, \ldots, x_{j,n}^{(i)}]$ represents a sequence of tokens from the j-th channel for the i-th example. The MGLM models a joint generative distribution over multiple channels: $p(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ using all possible factorizations.

2.1 Joint Probability Factorization

Given multiple sequences, each from different channels, there are many possible ways to factorize the joint probability over the channels. One approach to treat the channels as a *sequence* of channels, and use an autoregressive left-to-right model over the sequence of channels:

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_k) = p(\mathbf{x}_1) \prod_i p(\mathbf{x}_i | \mathbf{x}_1, \ldots, \mathbf{x}_{i-1})$$

Within each channel, the token sequence probability can also be modeled autoregressively:

$$p(\mathbf{x}_i | \mathbf{x}_1, \ldots \mathbf{x}_{i-1}) = \prod_t p(x_{i,t} | \mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_{i,<t})$$

This approach assumes: (1) a particular ordering over the channels; (2) the completion sequences from previous channels before generating the next channel's sequence. These assumptions are valid in some applications. For example, bilingual machine translation is a special case where $k = 2$, the channels are languages, and the source and target languages dictate the ordering over the channels and its token sequences.

In MGLM, we instead consider a more general approach, wherein we marginalize over all possible factorization order. Let z represent the permutation of indices $\{1, \ldots, N\}$ where N is the total number of tokens summed across all the channels. The joint probability is marginalized over z:

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_k) = \sum_{z \in S_N} p(z) p(\mathbf{x}_1, \ldots, \mathbf{x}_k | z), \quad (1)$$

Where $p(z)$ denotes a prior over the different possible permutation, which can be uniform or a balanced binary tree prior (Stern et al., 2019). Unfortunately, computing the exact log-likelihood in Eqn. 1 is intractable due to marginalization over all permutation order z. In practice, we optimize its lower bound via Jensen's inequality:

$$\log p(\mathbf{x}_1, \ldots, \mathbf{x}_k)$$
$$= \log \sum_{z \in S_N} p(z) p(\mathbf{x}_1, \ldots, \mathbf{x}_k \mid z) \quad (2)$$
$$\geq \sum_{z \in S_N} p(z) \log p(\mathbf{x}_1, \ldots, \mathbf{x}_k \mid z) =: \mathcal{L}(\{\mathbf{x}_i\}^k)$$
$$(3)$$

2.2 Model Architecture

One natural class of models for MGLM is the insertion-based Transformer (Stern et al., 2019; Welleck et al., 2019; Gu et al., 2019), which considers arbitrary factorization of the output sequence by using insertion operation, predicting both (1) content token $c \in \mathcal{C}$ from the vocabulary, and (2) location $l \leq t$ to insert, relative to (e.g. to the left of) the current partial output $\hat{\mathbf{y}}_t$:

$$p(c, l | \mathbf{x}, \hat{\mathbf{y}}_t) = \text{InsertionTransformer}(\mathbf{x}, \hat{\mathbf{y}}_t) \quad (4)$$
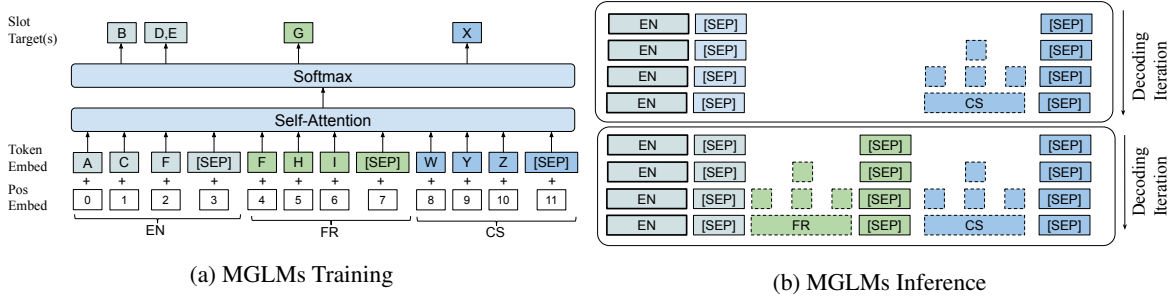
(a) MGLMs Training (b) MGLMs Inference

Figure 2: (a) An example multichannel modeling over 3 languages (English, French, Czech), where the model predicts the missing tokens at each location across multiple channels. (b) During inference, MGLM can generate output sequence for a single target language channel (top) or multiple language channels in parallel (bottom), conditioning on source channel sentence, and partial translations of multiple language channels.

The (content, location) distribution is factorized as $p(c, l) = p(c|l)p(l)$, where $p(c|l)$ is the standard Transformer softmax over vocabulary, and $p(l)$ is the softmax over the locations. KERMIT (Chan et al., 2019) further simplified the Insertion Transformer model by removing the encoder and only having a decoder stack (Vaswani et al., 2017), by concatenating the original input and output sequence as one single sequence and optimizing over all possible factorizations. Consequently, KERMIT is able to model the joint $p(\mathbf{x}, \mathbf{y})$, conditionals $p(\mathbf{x} \mid \mathbf{y})$, $p(\mathbf{y} \mid \mathbf{x})$, as well as the marginals $p(\mathbf{x}), p(\mathbf{y})$. We extend KERMIT to consider using a Transformer decoder for modeling the joint probability over $k > 2$ channels.

2.3 Training

Without the loss of generality, we denote $\mathbf{x} = [\mathbf{x}_1, \ldots, \mathbf{x}_k]$ as the concatenation of the k sequences[1]. With the insertion framework, the loss function Eqn. (3) can be simplified by changing the summation and careful decomposition of the permutation, leading to:

$$\mathcal{L}(\mathbf{x}) = \sum_{z \in S_N} p(z) \log \prod_{i=1}^{N} p((c_i^z, l_i^z) \mid \mathbf{x}_{1:i-1}^{z,i-1})$$

$$= \sum_{i=1}^{N} \sum_{z_{1:i-1}} p(z_{1:i-1}) \sum_{z_i} p(z_i \mid z_{1:i-1})$$
$$\log p((c_i^z, l_i^z) \mid \mathbf{x}_{1:i-1}^{z,i-1})$$

We illustrate an example data input consisting of 3 channels in Figure 2a. We concatenate the sequences together from all channels for each example, separated by a SEP token. Even with a shared

---
[1]The set of permutation $z \in S_N$ includes different order of channels as well

vocabulary, each channel results in a different token embedding, via addition of a channel-specific (learnable) embedding, or simply having a separately learned token embedding per channel. After passing through the dense self-attention layers as per Transformer architecture, the contextualized representation at each output time step predicts the possible tokens to be inserted to the left of the current input token. For a uniform prior $p(z)$, the target tokens at each slot are weighted equally.

2.4 Inference

At inference (generation) time, we can generate unconditionally by seeding the canvas with the [SEP] token and predicting the first actual token or provide as much, or as little, partial/complete sequence in each channel. Each output token is chosen via sampling or greedily choosing a single (content, location) with maximum probability in the partial canvas $\hat{\mathbf{x}}_t$:

$$(\hat{c}, \hat{l}) = \underset{c,l}{\operatorname{argmax}} \, p(c, l|\hat{\mathbf{x}}_t), \quad (5)$$

or inserted in all available insertion slots at once, in parallel:

$$\hat{c}_l = \underset{c}{\operatorname{argmax}} \, p(c \mid l, \hat{\mathbf{x}}_t), \quad (6)$$

Figure 2b shows two example decoding inference: a single target language channel (top), or multiple target language channels in parallel (bottom). Note that for both cases, each channel inserts in all available slots.

3 Related Work

MGLM was inspired and influenced by prior work on conditional and unconditional language modeling. Insertion Transformer (Stern et al., 2019) and

XLNet (Yang et al., 2019) also marginalize over all possible factorizations. However, their work is focused on the conditional distribution $p(\mathbf{y}|\mathbf{x})$, and they do not marginalize over all possible factorizations of the joint distribution. MGLM can be viewed as an extension and generalization of KERMIT (Chan et al., 2019). KERMIT is a generative joint distribution model that also learns all possible factorizations. However, KERMIT is constrained to two languages, while MGLM is a generative joint distribution model across any/all languages/text while learning all possible factorizations of the joint distribution.

MGLM follows from prior works on cross-lingual language models, which aim to learn shared representation across languages. XLM (Conneau et al., 2019) is closely related to our work and also concatenate source and target sequences from different languages; however, their work is limited to bilingual concatenation, is not fully generative, and requires length conditioning. MGLM is not limited to two languages and generalizes to multiple channels/languages, is fully generative, and our insertion-based approach (as opposed to masking-based approach) does not require length conditioning. Multilingual Neural Language Model (Wada and Iwata, 2018) uses a shared encoder and language-dependent decoders to generate word embeddings and evaluate word alignment tasks. In contrast, our work unifies the neural architecture with a straightforward stack of self-attention layers. Finally, Dong et al. (2015) explored multi-task learning for machine translation with an autoregressive network. The key difference between our work and other prior work on multi-target or multi-task learning is that MGLM models all possible factorizations of the joint distribution across all channels, instead of just the left-to-right factorization. This difference licenses MGLM to perform any form of sampling (conditional, unconditional, partially-conditional) without any rigid left-to-right restrictions.

Evaluation of text generative models remain a challenge (Liu et al., 2016; Novikova et al., 2017). Quality versus diversity plots have been used to compare the trade-off at different output softmax temperatures, as such in Stochastic Beam Search (Kool et al., 2019), which used a simpler n-gram diversity instead of Self-BLEU (Zhu et al., 2018). However, we are the first to characterize the Q-D behaviour of insertion based models versus existing left-to-right language models. Other metrics summarize the quality and diversity trade-off as a single number, such as Fréchet BERT Distance (Montahaei et al., 2019) inspired by the FID score (Heusel et al., 2017) used in computer vision, or take into account human evaluation (Hashimoto et al., 2019).

4 Experiments

We experiment on a multilingual dataset to demonstrate that we can learn MGLM. We perform both qualitative and quantitative experiments. We highlight the model's capabilities ranging from conditional generation (i.e., machine translation) to unconditional sampling of the joint distribution over multiple languages.

We experiment on the Multi30k[2] (Elliott et al., 2016b, 2017; Barrault et al., 2018), a multilingual dataset which consists of 29,000 parallel training sentences in English (EN), French (FR), Czech (CS), and German (DE) sentences. We use Multi30k because multiple high-quality channels (multilingual translations) are readily available to highlight our framework. We implement MGLM as a base Transformer decoder, without any causal masking, with 6 hidden layers and 1024 dimensional hidden representation. We concatenate all 4 language raw text training examples and use SentencePiece (Kudo and Richardson, 2018) to learn a universal subword unigram (Kudo, 2018) tokenizer with a shared 32K vocabulary size. We follow a similar training set up to BERT (Devlin et al., 2019), using Adam (Kingma and Ba, 2015) optimizer with a learning rate of 1e-4, warm up over the first 10% of the total training iterations varying between 10k to 50k iterations. We can train 3 different variants of MGLM by altering the sampling ratio of training data seen by the model:

1. **Bilingual** (e.g., EN → FR). We give the model a fully observed source (e.g., EN), and ask the model to infill the target (e.g., FR).

2. **Multi-target** (e.g., any 1 → Rest). We give the model a fully observed source (e.g., EN), and ask the model to infill the rest of the targets (e.g., DE, FR, CS).

3. **Joint**. We ask the model to infill all the targets, consequently we learn a joint distribution over all the languages $p(\text{en}, \text{fr}, \text{de}, \text{cs})$.

---

[2] https://github.com/multi30k/dataset

Model	Inference	Test2016	Test2017	MSCOCO
Bilingual (EN $\rightarrow$ DE)	EN $\rightarrow$ DE	36.14	28.32	24.15
Bilingual (EN $\leftrightarrow$ DE)	EN $\rightarrow$ DE	**37.08**	28.69	26.11
Multi-target (EN $\rightarrow$ Rest)	EN $\rightarrow$ DE	36.83	28.35	25.14
	EN $\rightarrow$ FR,CS,**DE**	35.41	**29.69**	25.64
Multi-target (Any $\rightarrow$ Rest)	EN $\rightarrow$ DE	36.63	28.37	**26.98**
	EN $\rightarrow$ FR,CS,**DE**	36.51	28.53	25.84
Joint ($p(EN, FR, CS, DE)$)	EN $\rightarrow$ DE	33.06	23.42	21.39
	EN $\rightarrow$ FR,CS,**DE**	32.53	23.78	20.97

Table 1: Multi30k English $\rightarrow$ German test BLEU. Higher is better.

EN Input: A man sits on a bench holding his dog and looking at the water.
Parallel Decode:

FR: _Un _homme _est _assis _sur _un _banc , _ten ant _son _chien _et _regardant _l ' eau . [SEP]
CS: _Muž _sedí _na _lavičce _a _drží _své ho _psa _a _dívá _se _na _vodu . [SEP]
DE: _Ein _Mann _sitzt _auf _einer _Bank _und _hält seine n _Hund _und _schaut _auf _das _Wasser . [SEP]

FR: _Un _homme _est _assis _sur _un _banc , _ten ant _son _chien _et _regardant _l ' eau . [SEP]
CS: _Muž _sedí _na _lavičce _a _drží _své ho _psa _a _dívá _se _na _vodu . [SEP]
DE: _Ein _Mann _sitzt _auf _einer _Bank _und _hält _seine n _Hund _und _schaut _auf _das _Wasser . [SEP]

FR: _Un _homme _est _assis _sur _un _banc , _ten ant _son _chien _et _regardant _l ' eau . [SEP]
CS: _Muž _sedí _na _lavičce _a _drží _své ho _psa _a _dívá _se _na _vodu . [SEP]
DE: _Ein _Mann _sitzt _auf _einer _Bank _und _hält _seine n _Hund _und _schaut _auf _das _Wasser . [SEP]

FR: _Un _homme _est _assis _sur _un _banc , _ten ant _son _chien _et _regardant _l ' eau . [SEP]
CS: _Muž _sedí _na _lavičce _a _drží _své ho _psa _a _dívá _se _na _vodu . [SEP]
DE: _Ein _Mann _sitzt _auf _einer _Bank _und _hält _seine n _Hund _und _schaut _auf _das _Wasser . [SEP]

FR: _Un _homme _est _assis _sur _un _banc , _ten ant _son _chien _et _regardant _l ' eau . [SEP]
CS: _Muž _sedí _na _lavičce _a _drží _své ho _psa _a _dívá _se _na _vodu . [SEP]
DE: _Ein _Mann _sitzt _auf _einer _Bank _und _hält _seine n _Hund _und _schaut _auf _das _Wasser . [SEP]

Figure 3: Example parallel greedy decode using the Multi-target (Any $\rightarrow$ Rest) KERMIT model, starting with an English sentence. Blue underlined tokens are the inserted tokens at each iteration, and the gray tokens are the final output tokens that have not been generated yet. The three target languages are generated together in parallel.

4.1 Translation Performance

The goal of MGLM is not conditional generation (i.e., machine translation), but nevertheless, we demonstrate its ability to do conditional generation in this section. We report the BLEU scores (Papineni et al., 2002) on the three test sets: test 2016 Flickr, test 2017 Flickr, test 2017 MSCOCO, for different English $\rightarrow$ {German, French, Czech} translations. We use parallel greedy decoding (Stern et al., 2019; Chan et al., 2019), i.e. inserting to all incomplete slots. Table 1 summarizes the results for English to German. Additional results for English to French, English to Czech, and German to English are shown in Appendix A.3. We observe that the Multi-target models performed similar to or slightly better than the bilingual models trained only on a single language pair. This is particularly useful when multiple machine translation targets are desired. We now only need one MGLM, which is competitive to the bidirectional expert models. This implies we only need 1 model for inference over multiple languages, instead of N models (i.e., saving substantial memory).

We also observe the full generative joint model has a BLEU gap compared to the bilingual baseline, consistent with the findings in Chan et al. (2019). We hypothesize this is due to the joint distribution being a more challenging task. We further hypothesize that the joint model needs to fantasize additional details when conditioning on the partial sequence in each channel during training. This results in fantasizing additional details not present in the source sentence during translation tasks.

4.2 Parallel Greedy Decoding: Parallel in Target Languages

As alluded conceptually in Figure 2 and in the previous section, our KERMIT-based MGLM is also able to perform parallel greedy decoding that is also *parallel in the number of target languages*. We illustrate this process in Figure 3. By starting with K initial [SEP] tokens for K target output

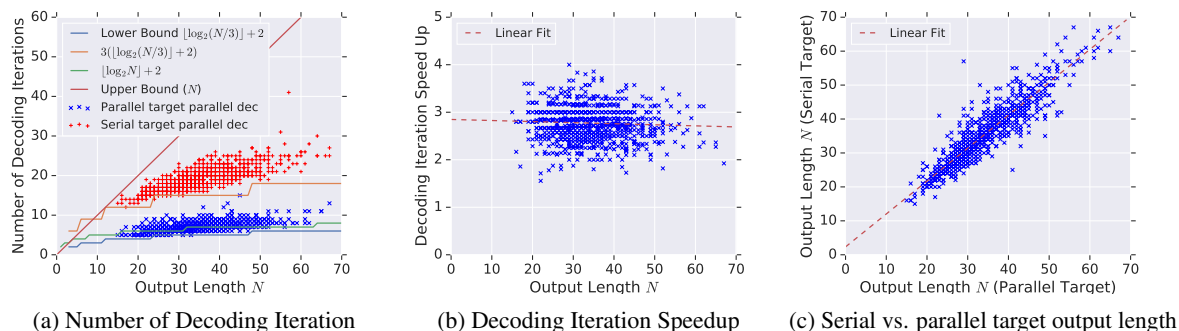| (a) Number of Decoding Iteration | (b) Decoding Iteration Speedup | (c) Serial vs. parallel target output length |

Figure 4: (a) The number of decoding iterations vs. the output length when decoding each target language serially vs. in parallel, compared to various logarithmic bounds. We have shown that the model can achieve close to the theoretical lower bound $\lfloor \log_2(N/k) \rfloor + 2$ where the number of target languages $k = 3$. (b) Relative wall-clock speed up when using the parallel target languages decoding vs. serial, achieving slightly under 3 times speed up. (c) Total output length for the 3 target languages when using serial vs. parallel target language generation. While not identical, we observe a linear relationship between the output length using the two different modes

languages, MGLM can decode K target languages that have at most n output tokens per language in $\mathcal{O}(\log n)$, i.e. constant in the number of target languages.

We investigate the relative speed up in generating multiple target language outputs in parallel versus generating the targets in series, in terms of wall-clock time and the number of decoding iterations. In Figure 4a, we plot the number of decoding iterations taken versus the total output length N for each sentence in the `test 2016 Flickr` test set, using the Joint MGLM model when decoding from a single source language to 3 target languages: English → {French, German, Czech}. When performing serial target decoding, we only output the target conditioned on English, i.e., English → French, English → German, English → Czech. We also plot several theoretical bounds: (1) upper bound (N) when decoding entirely serially, (2) lower bound $3(\lfloor \log_2(N/3) \rfloor + 2)$ when decoding 3 languages serially but parallel within each language, (3) lower bound $\lfloor \log_2(N/3) \rfloor + 2$, when decoding the 3 target languages in parallel and parallel within each language, and (4) $\lfloor \log_2(N) \rfloor + 2$, if we decode the entire output in parallel as a single sequence. We observe that our model can meet the lower bound several times and often decode below the fourth $\lfloor \log_2(N) \rfloor + 2$ bound. Figure 4b compares the wall-clock speed up when decoding targets in parallel vs. in series, with a linear regression line plotted. Our model achieving almost 3 times speedup in wall-clock speed. The parallel targets decoding is bottlenecked by the target language with the longest output sequence. Figure 4c

compares the total output length when decoding the targets in series versus in parallel. We observe that there is a linear relationship between the output lengths using the two modes.

4.3 Conditional Bilingual Generation: Quality-Diversity Trade-off

We first evaluated the models on *conditional* generation task by sampling bilingual translations (1 source, 1 target language) for each of the 12 language pair directions. We sample the token and location $(c, l) \sim p(c, l | x, \hat{y})$ from the partial canvas at each iteration, generating 100 hypothesis translations per source sentence, at softmax temperature $\tau = 0.1, 0.5, 1.0$. At each temperature and model, we computed the *quality* of the generated samples by computing the BLEU score between the reference translation and the samples, and the *diversity* by computing the pairwise BLEU between the 100 samples per source, also known as Self-BLEU (Zhu et al., 2018). Lower Self-BLEU indicates the higher diversity as there is less overlap between the samples.

Figure 5 illustrates the Quality-Diversity trade-off for the three models for different translation pairs involving English as one of the languages. The top right portion of the graph is the ideal area. We observed that the Multitarget model outperformed the Bilingual model at a lower temperature (both higher quality and diversity), and at a higher temperature, slightly above or below in quality but still higher diversity. Note that only one single Multitarget model was used for all language pair at inference time, while each bilingual model was
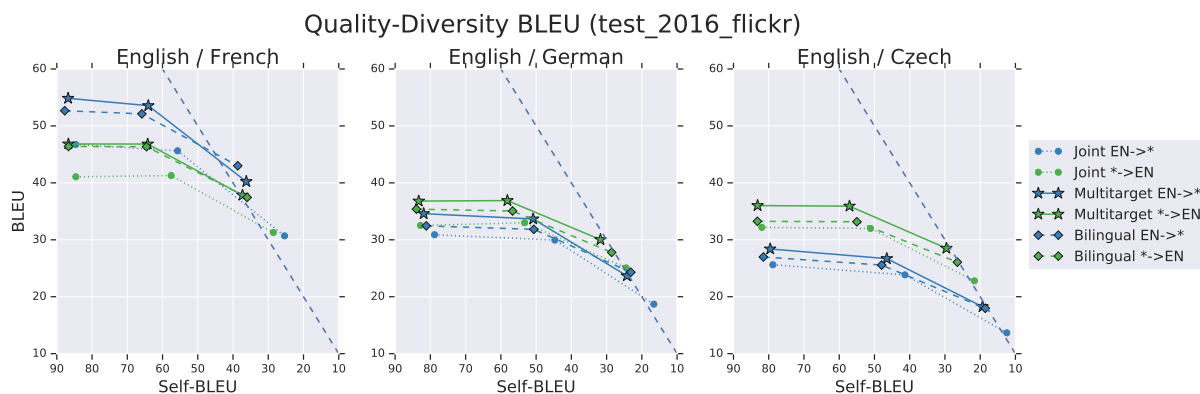
Figure 5: Quality-Diversity BLEU curve for several MGLM models (bilingual, multitarget, joint) on the Multi30k `text 2016 Flickr` test set. The dotted diagonal line signifies BLEU equals Self-BLEU. Points indicate different temperatures, from 0.1 (low diversity, left in the graph) to 1.0 (high diversity, right in the graph)

....
English Groundtruth: A young boy, wearing a chef's hat and apron, is cutting sausages in a kitchen.
French Groundtruth: Un jeune garçon, portant une toque et un tablier, coupe des saucisses dans une cuisine.
German Groundtruth: Ein kleiner Junge mit Kochmütze und Schürze schneidet in einer Küche Würstchen.
....
English Seed: A young boy,
French Seed: portant une toque et un tablier,
German Seed: chneidet in einer Küche Würstchen.
....
English: A young boy , wearing a hat , and an apron grilling hotdogs in the kitchen.
French: Un jeune garçon portant une toque et un tablier, faisant cuire du citron et des hotdogs dans la cuisine.
German: Ein junger Mann trägt eine Mütze und schneidet in einer Küche Würstchen.

English: A young boy , wearing a hat and a apron, is in a kitchen , cutting with various foods on it.
French: Un jeune garçon, portant une toque et un tablier, est dans une cuisine en projetant des poêles de la nourriture.
German: Ein kleiner Junge mit Hut und Schürze schneidet in einer Küche Würstchen.

English: A young boy, wearing an orange hat and apron, puts barbecue chicken in a kitchen.
French: Un jeune garçon, portant une toque et un tablier, coupant du poulet dans une cuisine.
German: Ein kleiner Junge in einer weißen Mütze und mit Schürze schneidet in einer Küche Würstchen glas .

English: A young boy, wearing a blue hat and apron, is cooking meat in a kitchen.
French: Un petit garçon, portant une toque et un tablier, fait la cuisine dans une cuisine.
German: Ein kleiner Junge mit blauer Mütze und schneidet in einer Küche Würstchen.

Figure 6: Example partially conditional generation samples. The seed text is shown in gray, with several different in-filling samples from the model in black. The samples show reasonable consistency and diversity.

different for each language pair curve. Therefore, a single Multitarget MGLM model could outperform specialized bilingual KERMIT models.

4.4 Partial Conditioning Multilingual Generation

We demonstrate our model's ability to generate in-filling for partial conditioning over the multiple channels. To be explicit, we seed each channel with a few (different) words, and sample from the model. We ask the model what text completions would best fit under the model's posterior. Figure 6 highlights several examples for (English, French, German) sentence completion. We took an example from the `test 2016 Flickr` test set and split it into 3 chunks–beginning in English, middle

in French, and ending in German–and sample completion. The model can generate a set of diverse, coherent examples (Figure 6).

4.5 Unconditional Multilingual Generation

We then evaluated the models on *unconditional* multilingual generation task to generate a sentence each in all 4 languages such that they correspond to each other. For the Joint model, we perform 3 types of sampling: (1) unrestricted, (2) chain, and (3) common cause. For unrestricted, we sampled one (token, location) at each iteration starting from an empty canvas, allowing the model to insert a token in any language until all slots were marked as completed. In the chain generation, we first restrict to generating English sentence one token at

Model	Language	Generated Sentences
Joint	English	A young man in a blue jacket walking up a mountain.
	French	Un jeune homme en veste bleue descendant une paroi rocheuse en horu.
	German	Ein junger Mann in einer blauen Jacke klettert eine Felswand hoch.
	Czech	Mladý muž v modré bundě stoupá po horách.
		≈*"Young men in blue jackets ascend and climb mountains."* ✓
Biling.	English	Two small white dogs are holding the duck in a fenced yard.
	French	Deux petits chiens blancs tenant un canard dans une cour clôturée.
	German	Zwei kleine weiße Hunde halten eine gelbe Ente in einem eingezäunten Hof.
	Czech	Dva malí chlapci drží žlutou panou venku u žlutého oploceném nádvoří.
		≈*"Two little boys holding a yellow gentleman outside by a yellow fenced courtyard."* ✗

Figure 7: Example unconditional text generation samples from the Joint (top) and chain of Bilingual model (bottom). The Joint model generates one long sequence, and we split them into the four sentences in each language. In contrast, Bilingual generates a complete sentence in each language conditioned on the previous sentence above.
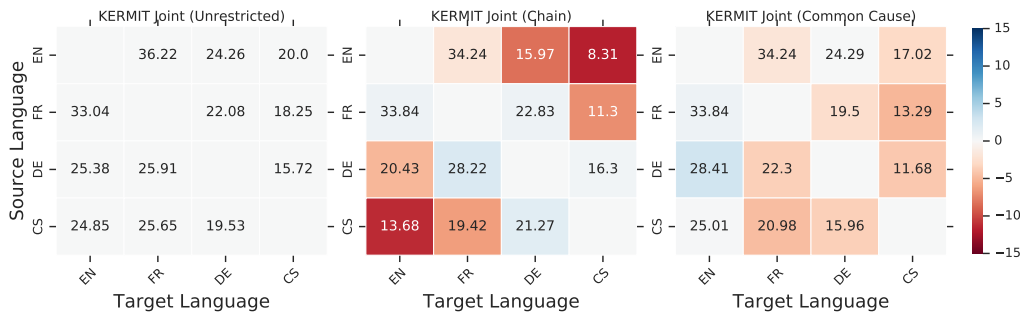

Figure 8: Pseudo-Target BLEU for self-consistency for unconditional multilingual generation. Colour shading indicates the difference compared to the Joint model (unrestricted) generation.

a time, then sampled French, German, and Czech in order, conditioned on the last sentence in the previous language. For common cause, we reuse the same English and French sampled sentences and generate the German and Czech conditioned on the English sentence (i.e., 3 languages are all conditioned on English).

Given these sets of sentences in 4 languages, for each pair of language direction, we computed a *pseudo* target by using a separately trained (on Multi30k) vanilla Transformer (Vaswani et al., 2017) and performed beam search (size 5) to translate the chosen source language sample. Figure 8 visualizes the pseudo target BLEU score for different source-target language pairs when comparing the Joint model under different types of sampling. The shaded colour represents the difference between the current sampling scheme versus the unrestricted reference. We observe that letting the model sample in unrestricted order was better than either the chain or the common cause sampling.

5 Conclusion and Future Work

In this paper, we presented the Multichannel Generative Language Model (MGLM). MGLM is a gen-

erative joint distribution model that marginalizes over all possible factorizations within and across channels. MGLM endows flexible inference, including unconditional, conditional, and partially observed generation. We experimented with those inference modes using the Multi30K dataset containing English, French, Czech, and German. We provide qualitative samples sampled unconditionally from the generative joint distribution. We also quantitatively analyze the quality-diversity trade-offs and find MGLM outperform traditional bilingual discriminative models.

Our work focused on a specific instantiation of channels as languages. However, MGLM is not limited to only languages and can generalize to other notions of channels. In future work, we will consider other textual channels, such as paraphrases, premises and hypotheses, questions and answers, and multimodal channels, such as images. Another direction can investigate scaling MGLM to dozens/hundreds of channels. Fully generative models still often lag behind purely discriminative counterparts in performance, but we hope our work motivates future research on building generative joint distribution models of the world.

References

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323.

William Chan, Nikita Kitaev, Kelvin Guu, Mitchell Stern, and Jakob Uszkoreit. 2019. KERMIT: Generative Insertion-Based Modeling for Sequences. In *arXiv*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-Task Learning for Multiple Language Translation. In *ACL*.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016a. Multi30K: Multilingual English-German Image Description. In *ACL: Workshop on Vision and Language*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016b. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

Jiatao Gu, Qi Liu, and Kyunghyun Cho. 2019. Insertion-based Decoding with Automatically Inferred Generation Order. In *arXiv*.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *arXiv preprint arXiv:1904.02792*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637.

Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Wouter Kool, Herke Van Hoof, and Max Welling. 2019. Stochastic beams and where to find them: The Gumbel-top-k trick for sampling sequences without replacement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3499–3508, Long Beach, California, USA. PMLR.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Ehsan Montahaei, Danial Alihosseini, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. *arXiv preprint arXiv:1904.03971*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. *arXiv preprint arXiv:1707.06875*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Mitchell Stern, William Chan, Jamie Kiros, and Jakob Uszkoreit. 2019. Insertion Transformer: Flexible Sequence Generation via Insertion Operations. In *ICML*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *NIPS*.

Takashi Wada and Tomoharu Iwata. 2018. Unsupervised Cross-lingual Word Embedding by Multilingual Neural Language Models. In *arXiv*.

Sean Welleck, Kiante Brantley, Hal Daume, and Kyunghyun Cho. 2019. Non-Monotonic Sequential Text Generation. In *ICML*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100. ACM.

A Appendices

A.1 Multi30k Dataset Description

The statistics of the Multi30K dataset (Task 1) are summarized in Table 2. The average number of words across training, validation, and 2016 test for English is 11.9, and for German is 11.1 (Elliott et al., 2016a). Since we use SentencePiece (Kudo and Richardson, 2018), MGLM sees more number of tokens per sentence on average.

Subset	Number of Sentences
Training	29,000
Validation	1,014
Test 2016 Flickr	1,000
Test 2017 Flickr	1,000
test 2017 MSCOCO	461

Table 2: Multi30k English → Czech test BLEU.

A.2 Additional Quality-Diversity Curves For Conditional Generation

We include additional Quality-Diversity Curves For Conditional Generation: Figure 9 for the `test 2017 Flickr`, and Figure 10 for the `test 2017 MSCOCO`.

A.3 Additional Multi30K Translation Results

We include additional Multi30K Translation Results: Table 3 for English to French, Table 4 for English to Czech, and Table 5 for German to English.

A.4 Unconditional Sampling Generation

Figure 11 illustrates the serial sampling (one token at a time) from the joint model, every 20 timesteps.
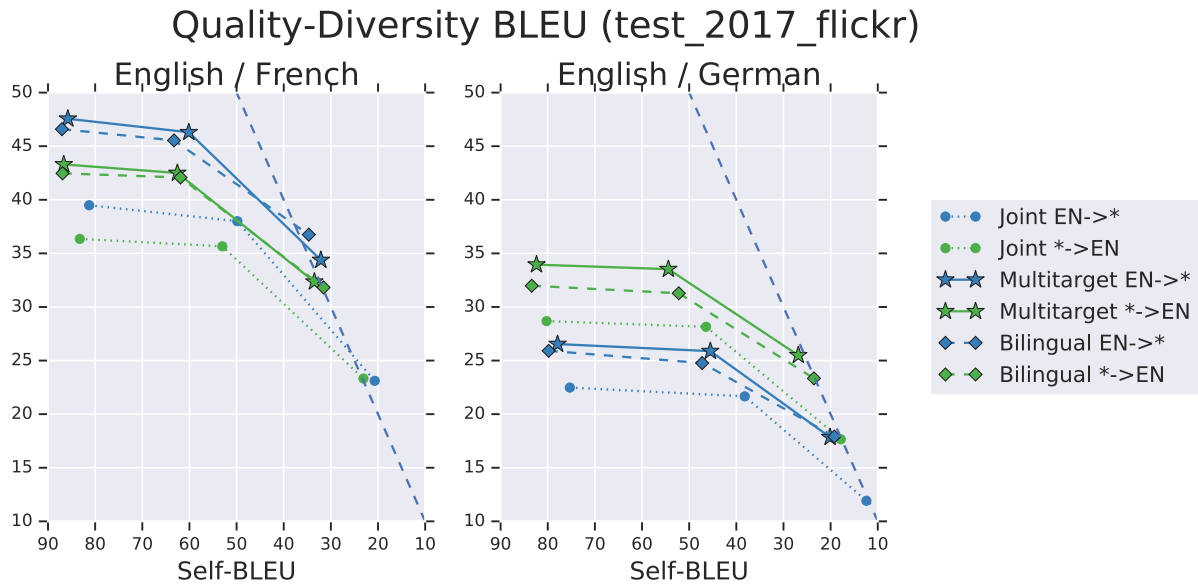
Quality-Diversity BLEU (test_2017_flickr)

Figure 9: Quality-Diversity BLEU curve for several MGLM models (bilingual, multitarget, joint) on the Multi30k `text 2017 Flickr` test set. The dotted diagonal line signifies BLEU equals Self-BLEU. Points indicate different temperatures, from 0.1 (low diversity, left in the graph) to 1.0 (high diversity, right in the graph)

Quality-Diversity BLEU (test_2017_mscoco)

Figure 10: Quality-Diversity BLEU curve for several MGLM models (bilingual, multitarget, joint) on the Multi30k `text 2017 MSCOCO` test set. The dotted diagonal line signifies BLEU equals Self-BLEU. Points indicate different temperatures, from 0.1 (low diversity, left in the graph) to 1.0 (high diversity, right in the graph)
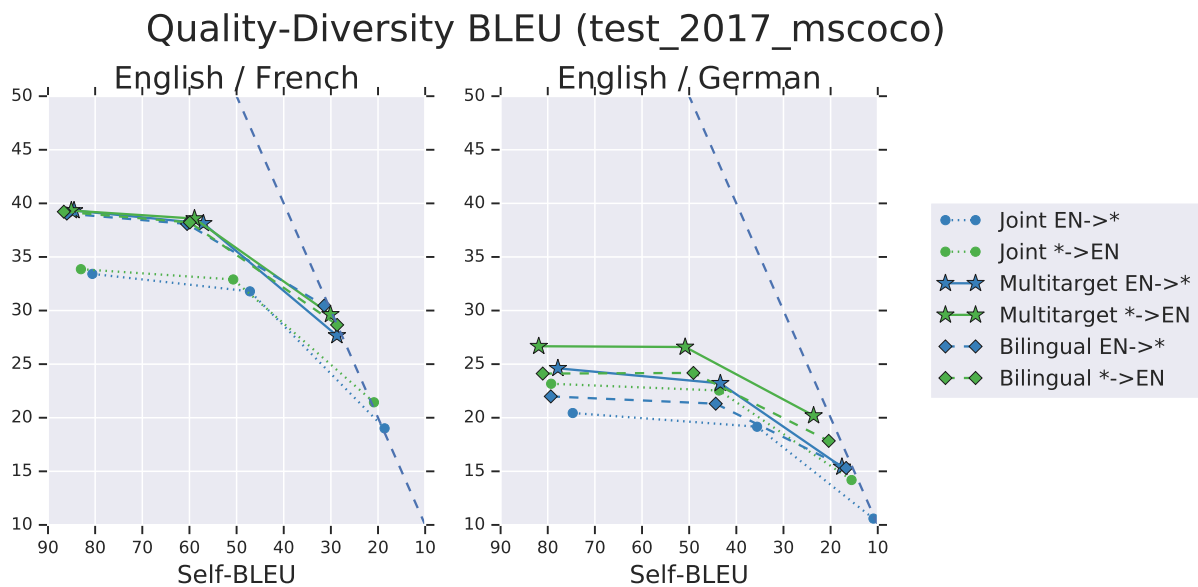
Model	Inference	Test2016	Test2017	MSCOCO
Bilingual (EN → FR)	EN → FR	58.80	50.35	**42.82**
Bilingual (EN ↔ FR)	EN → FR	**59.29**	**52.13**	42.17
Multi-target (EN → Rest)	EN → FR	58.08	50.39	42.19
	EN → **FR**,CS,DE	58.52	50.49	41.53
Multi-target (Any → Rest)	EN → FR	57.64	50.01	40.18
	EN → **FR**,CS,DE	57.35	48.13	39.98
Joint ($p(EN, FR, CS, DE)$)	EN → FR	50.87	40.69	33.93
	EN → **FR**,CS,DE	48.85	39.92	33.45

Table 3: Multi30k English → French test BLEU.

Model	Inference	Test2016
Bilingual (EN → CS)	EN → CS	28.58
Bilingual (EN ↔ CS)	EN → CS	29.03
Multi-target (EN → Rest)	EN → CS	**30.48**
	EN → FR,**CS**,DE	30.15
Multi-target (Any → Rest)	EN → CS	30.11
	EN → FR,**CS**,DE	30.11
Joint ($p(EN, FR, CS, DE)$)	EN → CS	26.45
	EN → FR,**CS**,DE	26.35

Table 4: Multi30k English → Czech test BLEU.

Model	Inference	Test2016	Test2017	MSCOCO
Bilingual (DE → EN)	DE → EN	39.40	34.90	27.75
Bilingual (EN ↔ DE)	DE → EN	40.52	35.66	28.61
Multi-target (DE → Rest)	DE → EN	**40.75**	36.38	**28.91**
	DE → **EN**, FR,CS	39.72	35.95	28.20
Multi-target (Any → Rest)	DE → EN	40.69	36.02	28.89
	DE → **EN**, FR,CS	39.97	**37.07**	28.62
Joint ($p(EN, FR, CS, DE)$)	DE → EN	38.44	30.82	25.46
	DE → **EN**, FR,CS	36.30	29.68	24.87

Table 5: Multi30k German → English test BLEU.

Iterations	Language	Generated Sentence from Joint Model
1	English French Czech German	 Mladý
20	English French Czech German	 descendant Mladý muž v modré bundě stoupá po Mann klettert.
40	English French Czech German	blue jacket walking up a mountain. veste descendant paroi rocheuse en Mladý muž v modré bundě stoupá po horách. Mann klettert.
60	English French Czech German	A man blue jacket walking up a mountain. veste bleue descendant une paroi rocheuse en horu. Mladý muž v modré bundě stoupá po horách. Mann einer blauen klettert eine hoch.
80	English French Czech German	A young man in blue jacket walking up a mountain. veste bleue descendant une paroi rocheuse en horu. Mladý muž v modré bundě stoupá po horách. Ein junger Mann in einer blauen Jacke klettert eine Felswand hoch.
96	English French Czech German	A young man in a blue jacket walking up a mountain. Un jeune homme en veste bleue descendant une paroi rocheuse en horu. Mladý muž v modré bundě stoupá po horách. Ein junger Mann in einer blauen Jacke klettert eine Felswand hoch.

Figure 11: Example of serial sampling unconditional text generation from the joint $p(EN, FR, CS, DE)$ model, over 96 insertion time steps. Note that the model generates one long sequence, and we split them into the resulting four sentences in each language here.