Learning to translate by learning to communicate

C.M. Downey[α*] **Xuhui Zhou** [β*] **Leo Z. Liu**[γ] **Shane Steinert-Threlkeld**[α]
[α]Department of Linguistics, University of Washington
[β]Language Technologies Institute, Carnegie Mellon University
[γ]Department of Computer Science, The University of Texas at Austin
{cmdowney,shanest}@uw.edu
zliu@cs.utexas.edu, xuhuiz@cs.cmu.edu

Abstract

We formulate and test a technique to use Emergent Communication (EC) with a pre-trained multilingual model to improve on modern Unsupervised NMT systems, especially for low-resource languages. It has been argued that the current dominant paradigm in NLP of pre-training on text-only corpora will not yield robust natural language understanding systems, and the need for grounded, goal-oriented, and interactive language learning has been highlighted. In our approach, we embed a multilingual model (mBART, Liu et al., 2020) into an EC image-reference game, in which the model is incentivized to use multilingual generations to accomplish a vision-grounded task. The hypothesis is that this will align multiple languages to a shared task space. We present two variants of EC Fine-Tuning (Steinert-Threlkeld et al., 2022), one of which outperforms a backtranslation-only baseline in all four languages investigated, including the low-resource language Nepali.

1 Introduction

While neural machine translation (NMT) systems are one of the great success stories of natural language processing (Sutskever et al., 2014; Bahdanau et al., 2015; Wu et al., 2016), typical methods rely on large quantities of *parallel text* (i.e. existing human translated texts) as gold data for supervised learning. These approaches are thus difficult to apply to low-resource languages, which lack large bodies of such data (Joshi et al., 2020). To extend this vital language technology to low-resource languages, many have focused on *Unsupervised NMT* (UNMT) — the task of building NMT systems without any parallel text (Artetxe et al., 2018; Lample et al., 2018a,c; Lample and Conneau, 2019; Conneau et al., 2020).

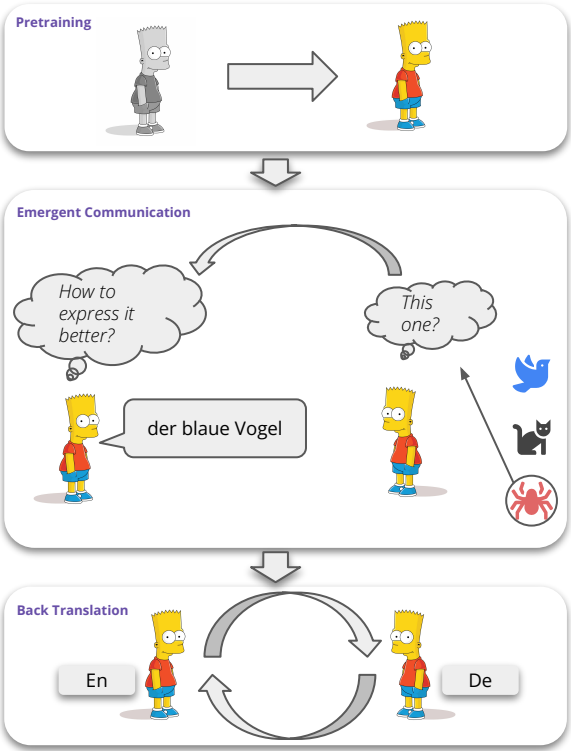

Figure 1: Illustration of our modeling process. For the *pre-training* stage, we use the off-the-shelf mBART (Lewis et al., 2020). We fine-tune the model for translation with Emergent Communication.

Typical approaches to UNMT rely on large pre-trained multilingual models (Lample and Conneau, 2019; Conneau et al., 2020; Liu et al., 2020; Song et al., 2019) and the method of *back-translation* (Sennrich et al., 2016b) to iteratively generate synthetic parallel text. These approaches, however, still rely on plain text information alone. For that reason, the resulting models are considered *ungrounded* (there is no link between the text and the external world). This may limit model abilities.

Despite NLP breakthroughs stemming from large-scale pre-training on raw text corpora with self-supervised learning (Howard and Ruder, 2018; Peters et al., 2018; Devlin et al., 2019; Liu et al., 2019; Conneau et al., 2020; Liu et al., 2020; Brown

*Equal contribution. We also include a detailed Author Contribution Statement at the end of the paper.

Proceedings of the The 3rd Workshop on Multi-lingual Representation Learning (MRL), pages 218–238

et al., 2020, i.a.), several recent results suggest limitations in model generalization (McCoy et al., 2019; Niven and Kao, 2019; Ettinger, 2020; Rogers et al., 2020, i.a.). More fundamentally, several have argued that pre-training on text alone will not deliver fully general and robust NLP systems.[1]

For example, using several detailed thought experiments, Bender and Koller (2020) argue that models trained on text alone will not, in principle, be able to recover either the conventional meaning of expressions or the communicative intent of an expression in context. Their arguments highlight the importance of the interaction between linguistic expressions and extra-linguistic communicative intents (e.g. acting in the world, executing programs).[2] Similarly, Bisk et al. (2020) articulate progressively broader *world scopes* in which language use is embedded, and argue that present pre-training methods work at a relatively limited scope. They too emphasize the importance of embodied interaction with the environment and with the social world for future NLP systems.[3]

In this paper, we propose to use methods from the field of *emergent communication* (EC) (Wagner et al., 2003; Skyrms, 2010; Lazaridou and Baroni, 2020) to improve UNMT systems. EC studies artificial agents communicating with each other to accomplish particular environmental goals. EC is a subfield of reinforcement learning, wherein language (i.e. the communcation protocol) is shaped by rewards determined by interacting with an external environment and with other agents. Typical work in this area starts from a *tabula rasa* and studies under what conditions —e.g. environments, tasks/goals, social settings—the resulting communication protocols among agents resembles human language, along axes like word length economy (Chaabouni et al., 2019a), word-order biases (Chaabouni et al., 2019b), and compositionality (Andreas, 2019; Chaabouni et al., 2020; Steinert-Threlkeld, 2020; Geffen Lan et al., 2020), among others (Mu and Goodman, 2021).

Our approach leverages the insight that people learn new languages by using them to do things (e.g. order food, buy train tickets); our machines should do the same. We improve upon a standard UNMT system by taking a large pre-trained multilingual model (mBART) and embedding it in an EC task, having it participate in goal-directed communication (in addition to back-translation). Communication should promote translation in the following way. Translation can be viewed as 'aligning' model representations for sentences in several languages. In the supervised case, parallel text instructs the model how to do this alignment. In the unsupervised case, through communication, each model aligns its language representations *with the same shared environment*, thereby promoting alignment between the languages themselves. This work is thus an instance of the wider framework of Emergent Communication Fine-tuning (EC-FT) (Steinert-Threlkeld et al., 2022).

In what remains, we describe our pipeline for EC fine-tuning (Section 2) and the experiments that we conduct to demonstrate its benefit for UNMT (Section 3), overview our experimental results, in which we show EC yields benefits for every language we study with particularly strong gains for the low-resource language Nepali (Section 4). We then study some manipulations on our training pipeline (Section 5) before discussing the implications of these experiments (Section 6), and situating them in the context of existing work (Section 7).

Our contributions are the following: (i) We demonstrate that EC-FT can be used to improve upon UNMT baselines. (ii) We give a proof-of-concept for the viability of using modern pre-trained language models in an EC scenario. (iii) We articulate a view for EC-FT as a generalized and parameterizable framework.

2 Methodology

As shown in Figure 1, the pipeline that we introduce here consists of three main phases: (1) Begin with a pre-trained multilingual model, which either already has an encoder and decoder, or from which this *seq2seq* stack can be initialized. (2) Conduct emergent-communication training using image and/or text embeddings (Figure 2). (3) Use iterative backtranslation (Sennrich et al., 2016a; Section 7) to tune the model for translation.[4]

[1] This is largely what (Linzen, 2020) calls the pre-training Agnostic Independently Distributed (PAID) evaluation paradigm. We discuss pre-training on multimodal (i.e. not text-only) data in § 7.

[2] See Merrill et al. (2021) for a formalization of argument in Bender and Koller (2020) about learning a programming language from form alone.

[3] As noted by Bender and Koller (2020), many of these arguments can be seen as detailed elaborations of the need for NLU systems to solve the *symbol grounding* problem (Harnad, 1990; Taddeo and Floridi, 2005).

[4] The code to run all experiments described here can be found at https://github.com/CLMBRs/communication-translation.

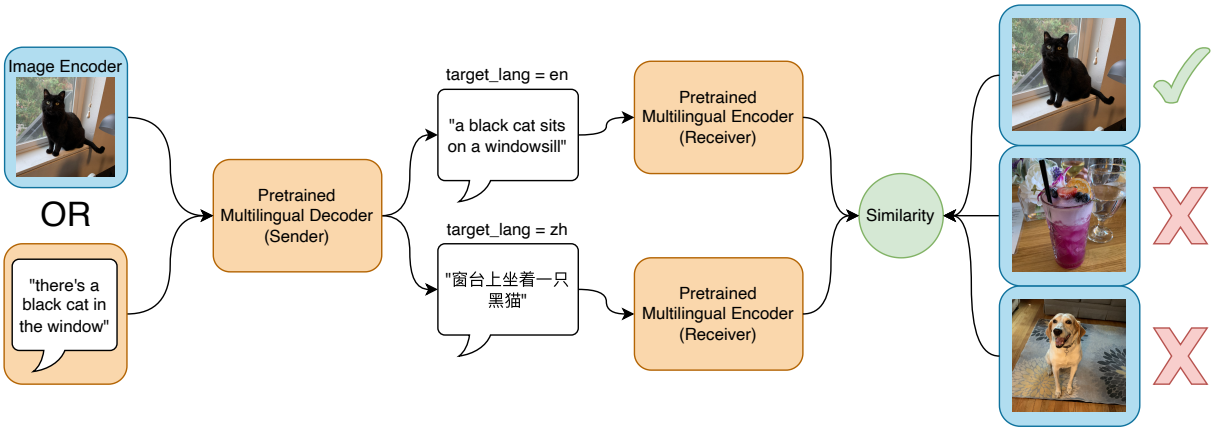

Figure 2: Emergent Communication Fine-Tuning: the task is a standard image reference game from the EC literature, but with the sender and receiver initialized from a pre-trained multilingual decoder and encoder. The communication language alternates between the two languages in the translation pair that is being fine-tuned.

For step (2), we test two versions of the EC fine-tuning task. In the first (I2I-EC), the EC step uses *only image embeddings*, and the model must select the original input image from among distractors, based on a text generation (akin to a caption). In the second (T2I-EC), the communication game involves gold captions, instead of only image features: based on a caption, the model must generate a translation of it, on the basis of which the original image must be selected from amongst distractors.

First, we introduce some notation. We use E_m and D_m for the multilingual encoder and decoder, respectively, which are parameterized by θ_E and θ_D. $\mathbf{x}_E \in \mathbb{R}^{N\times|V|}$ and $\mathbf{x}_D \in \mathbb{R}^{K\times|V|}$ are sequences of symbols of length N and K respectively. $E_m(\mathbf{x}_E;\theta_E) \in \mathbb{R}^{N\times d_m}$, where d_m is the model hidden dimension, is the encoder output. Similarly, the decoder output is $D_m(\mathbf{x}_D, \mathbf{e};\theta_D) \in \mathbb{R}^{K\times|V|}$, where $\mathbf{e} \in \mathbb{R}^{N\times d_m}$ is a set of vectors for cross-attention of the decoder.

This formulation of our pipeline leaves many concrete choices open. In the remainder of this section, we describe the specific implementation of this process used in our experiments.

2.1 Pipeline Components

Pre-trained Model We use mBART(-large) (Liu et al., 2020), which has demonstrated strong unsupervised translation performance in several languages. mBART employs *seq2seq* pre-training, encoding a "noised" input sequence and then reconstructing the original sequence with the decoder, over a collection of 25 languages. mBART's encoder-decoder architecture and corresponding seq2seq training make it a natural fit for our EC experiments, in which a multilingual decoder and encoder are used to send and receive natural language messages. We use $\theta_{E_{PT}}$ to denote the parameters of the pretrained encoder, and *mutatis mutandis* for the pretrained decoder.

Backtranslation Iterative backtranslation allows a model (usually pre-trained) to achieve some level of translation performance while only training on monolingual data (Section 7). Our baseline system is mBART fine-tuned with backtranslation only. In the EC-FT case, backtranslation is always performed last so that the model is tuned for translation immediately before it is evaluated.

Image-to-Image EC (I2I-EC) Our emergent communication framework consists of two main subtasks. First, an agent (the sender, a decoder) must take in an image encoding and produce a natural language description of it. The generation language may vary; there will be several in our experiments. Next, another agent (the receiver, an encoder) takes in the generated text and uses it to pick the described image from a set of distractors. In the EC literature, this is referred to as a standard image reference game (see Figure 2).[5]

Let $i \in \mathbb{R}^{d_i}$ be an image embedding (d_i is the dimension of these embeddings, which may come from a vision model). We also assume that we have

[5]The image reference game, as used in much of the EC literature, is very similar to? training an image caption module to produce discriminative captions via self-retrieval, as pursued in (Liu et al., 2018). They first train the text-to-image pipeline from gold captions, and then pursue training a caption generator via image selection both with and without supervision from gold captions. We thank an anonymous reviewer for calling our attention to this work.

a reshaper $R(i;\theta_R)$ which maps images to $\mathbb{R}^{d_m}$.

Because mBART is not natively multi-modal, some adaptations are made to allow it to generate a description of an image. In particular, the image embedding cannot simply be the first token to the sender since mBART reserves this for a special language identification token. Further, it is not obvious that a pre-trained transformer decoder's cross-attention can be "turned off" without effecting overall performance. For these reasons, we pass the image embedding into an "unroller" U (one auto-regressive transformer layer) to generate a sequence of embeddings $U(R(i);\theta_U) \in \mathbb{R}^{M\times d_m}$ where M is a hyperparameter. This sequence is then used as the keys and values in the sender's cross-attention.

We auto-regressively generate from the sender's distributions $S := D_m\left(\langle \text{LID}, T_{<K}\rangle, U(R(i))\right) \in \mathbb{R}^{K\times|V|}$, where LID is a language ID token and $T_{<K}$ is the prefix of text T generated at the previous time step. The sampling required for discrete generation is not differentiable, so we use the straight-through Gumbel-Softmax estimator (Jang et al., 2017; Maddison et al., 2017) with temperature $\tau = 1.0$. $T := \text{GS-ST}(S)$ is the sequence of one-hot vectors sampled in this way.

The receiver consumes this generated 'caption': $E_m(T) \in \mathbb{R}^{K\times d_m}$. To produce a single representation of the image, we use an 'aggregator' A which takes this sequence of representations and pools them into a single one $A(E_m(T);\theta_A) \in \mathbb{R}^{d_m}$.[6]

The score for each of the candidate images is the inverse of the mean squared error between the image and the receiver's final representation. The loss for the image selection task is then cross-entropy among the image candidates. This loss partially follows Lee et al. (2018), though they jointly train on supervised caption generation during EC.

Given the original image i, and a set $\{i_m\}_{m=1}^{M}$ of distractor images, let the image selection loss be

$$\ell_{\text{IS}}(i,\Theta) := -\log \text{softmax} \frac{1}{\|A(E_m(T)) - R(i)\|_2^2} \quad (1)$$

where $\Theta = \{\theta_D, \theta_E, \theta_R, \theta_A, \theta_U\}$ and the softmax is taken over the distractor images $\{R(i_m)\}$.

Finally, because EC can cause significant language drift (Lee et al., 2018, 2019; Lu et al., 2020; Lazaridou et al., 2020), we use KL regularization (Havrylov and Titov, 2017; Baziotis et al., 2020) to ensure that the sender's output distribution does not drift too far from the distribution of an auxiliary causal language model (CLM; this model is not trained as part of EC):[7]

$$\ell_{\text{KL}} := \frac{1}{K}\sum_k \text{KL}\left(S_k \;||\; D_{\text{CLM}}\left(\langle LID, T_{<k}\rangle\right)_k\right) \quad (2)$$

Combining equations (1) and (2) and averaging over iterations of the game, the final EC loss is

$$\mathcal{L}_{\text{EC}} := \mathbb{E}_i\left[\ell_{\text{IS}} + \lambda \ell_{\text{KL}}\right] \quad (3)$$

with λ a hyperparameter.

Text-to-Image EC (T2I-EC) The text-to-image EC task is identical to I2I-EC, except in what is presented to the sender via cross-attention. In T2I-EC, monolingual gold captions are used in the cross-attention for the emergent generation after being embedded by the encoder E_m.

In other words, given c_i as a caption for image i, T2I-EC still uses $\mathcal{L}_{\text{EC}}$ (equation (3)), but without the unroller for the sender. Now, we have $S = D_m\left(\langle \text{LID}, S_{<K}\rangle, E_m(c_i;\theta_E)\right)$.

As in I2I-EC, the image descriptions are generated in *either* the caption language (here, English) or another translation target language. Importantly, the emergent generation need not be identical to the gold caption. This is desirable, since there may be several valid paraphrases of a given translation/caption. Similarly, we only require gold captions in one language, not every language; for this reason, there is no implicitly parallel text data and so the translation task can still be considered unsupervised.

The motivation for this version of EC comes from the observation that the encodings used in the sender's cross-attention should be fairly similar to those generated by the model's encoder, since the model is being fine-tuned to be an encoder-decoder translation model. Generating into varying target languages incentivizes the model to use the same encodings for generating different languages, rather than copying the input text to the output. In contrast, there is no guarantee that the image encodings used in I2I-EC are at all similar to those produced by the model's encoder.

[6] Pilot experiments suggested that a small aggregator worked better than simply using mean pooling.

[7] We finetune the original mBART decoder as a CLM for this purpose; see the end of Appendix A for details.

Initial Supervision Because multilingual EC is a complicated task with sparse training signal, we first ground the agents in their visual sub-tasks independently of the combined communication task. We train the sender to produce gold-standard captions in a high-resource language (English in our experiments) while simultaneously training the receiver to pick out the correct image based on the gold-standard caption. Critically, this stage only assumes that you have gold-standard captions in *one* language. The model is never trained on gold captions in non-English languages. This step is conducted independently, before EC.

2.2 Data

Training We use two main sources of training data: monolingual corpora for backtranslation, and pairings of images and captions in a single high-resource language. We train translation systems between English and four other languages: Chinese (zh), German (de), Nepali (ne), and Sinhala (si).

Backtranslation creates synthetic translation pairs by generating sentences in the second language given natural sentences in the first. Following experiments using mBART for unsupervised translation (Liu et al., 2020), we use small portions of the Common Crawl 25 dataset, which is the pre-training data for mBART. In this way, no novel data is introduced to establish our UNMT baseline.

For the EC stage, the data required differs between I2I-EC and T2I-EC. The former requires only image embeddings. The latter requires paired images and captions, since the true caption is used to prompt the sender's generation. As mentioned, we assume that captions are *only available for one language*. Since English is in every translation pair, we use English captions. Our image-caption pairs come from the MS-COCO dataset (Lin et al., 2014), and our image embeddings are extracted from ResNet 50 (He et al., 2016b) (these are also used during the supervised captioning stage).

Validation and Test Translation validation and test sets are the only parallel data used in our experiments. For Nepali and Sinhala, we use the standard splits of the FLoRes evaluation datasets (Guzmán et al., 2019). For Chinese and German, we use the `newstest2018` and `newsdev2019` splits of the WMT'19 release as validation data (Barrault et al., 2019). For test data in these two languages, we sample 4096 examples from News Commentary v14 subset of the same release.

3 Experiments

We evaluate a UNMT baseline and our two proposed EC-FT pipelines on translation performance for each language pair. Checkpoints are picked by highest mean BLEU on the validation set. We first describe these models and then our evaluation. More extensive details can be found in Appendix A.

Baseline For our UNMT baseline, we start with mBART-25 and perform iterative backtranslation for 8192 steps in each direction. mBART employs language control tokens at the beginning of sequences, but it is *not* pre-trained to decode one language from another (Liu et al., 2020), which is a key feature of (back-)translation. To overcome the model's tendency to copy the input sequence to the output, we establish language-controlled generation using language control tokens and language masks (Liu et al., 2020). Concretely, we obtain token counts from the mBART training data, and these are used to create a logit mask, only allowing the model to generate tokens which make up the top p percent of the probability mass of the data in the given language. For the first 2048 backtranslation steps, we use a masking threshold of $p = 0.9$. After that, we raise the threshold to $p = 0.99$.

(I2I/T2I)-EC In both of our EC-FT models, we keep the total number of backtranslation steps the same (8192), and add 2048 steps each of supervised caption training and EC-FT. The language of generation can also be controlled during EC, so we use language-control tokens and a logit mask to ensure the sender generates in the specified language. The language of the emergent generation is selected uniformly at random per example.

Evaluation For our final evaluation, we report both BLEU and COMET (Rei et al., 2020) scores in both translation directions for each language pair. COMET provides the output of a regression model trained to predict the human direct-assessment translation quality score of a translation pair. Based on normalized quality scores, a COMET score of 0 means the translation is predicted to be of average quality. Postive scores indicate above-average quality, and vice-versa. We use the `wmt22-comet-da` model.

4 Results

Table 1 shows the results from our main experiment. Firstly, our UNMT baseline based on iterative back-

Model	Language	BLEU				COMET		
		en→X	X→en	mean	Δ	en→X	X→en	mean
baseline (mBART + BT)	zh	18.45	11.36	14.90	-	0.03	0.15	0.09
	de	19.06	25.73	22.39	-	0.20	0.38	0.29
	ne	2.14	5.07	3.60	-	-0.24	-0.34	-0.29
	si	1.18	4.73	2.95	-	-0.18	-0.28	-0.23
I2I-EC	zh	18.72	11.88	15.30	+3%	0.04	0.17	0.10
	de	18.26	25.60	21.93	-2%	0.20	0.40	**0.30**
	ne	1.51	5.34	3.43	-5%	-0.24	-0.31	-0.28
	si	0.01	0.08	0.04	-99%	-1.31	-1.05	-1.28
T2I-EC	zh	19.25	11.91	**15.58**	+5%	0.06	0.18	**0.12**
	de	18.64	26.20	**22.42**	+0.1%	0.19	0.41	**0.30**
	ne	2.36	5.92	**4.14**	+15%	-0.20	-0.27	**-0.24**
	si	1.29	4.76	**3.02**	+2%	-0.18	-0.27	**-0.22**

Table 1: Results of our main experiment. Values reported here are the maximum across 3 random seeds per row; see Appendix C for full variation. T2I-EC shows consistent improvement for each language in terms of both mean BLEU and COMET. Δ shows percent improvement over the baseline.

translation (BT) shows a marked decrease in performance from the two higher-resource languages (Chinese and German) to the two lower-resource languages (Nepali and Sinhala). This is expected since BT-based UNMT often requires a strong initialization (Lample et al., 2018c) and multilingual models (like mBART) do not perform as well for lower-resource languages (Wu and Dredze, 2020).

Our model fine-tuned with both backtranslation and I2I-EC remains close to or exceeds the baseline for the two higher-resource languages and Nepali but achieves very poor performance on Sinhala. It appears that EC provides a worse initialization for backtranslation for this language.

In contrast, our "text-to-image" variant of EC-FT (T2I-EC) yields the best performing model in terms of mean BLEU for all four of our languages. In particular, we see significant gains for both lower-resource languages. Most striking is the Nepali-English pair, which sees a +15% BLEU improvement over the baseline. While there are improvements in both directions, the Nepali→English direction has the largest gain. By contrast, Sinhala shows improvements in both directions, with the larger improvement in the to-Sinhala direction (partially due to a stronger baseline). The improvements are smallest for German, which is both very high-resource and the most similar to English of our languages. The COMET scores were broadly correlated with BLEU scores in all of our settings.

These results show that EC-FT of a pre-trained multilingual model can provide real improvement over a backtranslation-only baseline, giving proof-of-concept of communication for fine-tuning.

5 Manipulations

To better understand which components of the pipeline affect the results in T2I-EC, we conducted several follow-up experiments. For each manipulation, we looked at one high-resource language (German) and one low-resource langauge (Sinhala). See Appendix B for full methodological details.

Image Encoder To test the effect of the image encoder, we replaced the ResNet image encoder with the best performing one from CLIP (Radford et al., 2021). This image encoder is based on the Vision Transformer (Dosovitskiy et al., 2021) architecture and trained jointly with a text encoder via a contrastive loss to pair image encodings with caption encodings.

Initial Backtranslation Because the EC component of training is the first time that language ID codes are being used to generate text from the decoder with input other than representations of the same language from the encoder, we experimented with splitting the backtranslation training into two parts. Instead of doing all 8192 steps after EC, we did 2048 steps after image supervision but before EC, and the final 6144 steps after EC.

Interleaved Training Inspired by Lowe et al. (2020), who showed that inter-leaving EC with a supervised learning objective can improve EC results, we ran a version of our training pipeline where we alternated between EC and BT four times. The total number of training steps remained the same (2048 and 8192, respectively), but this was now done in 4 equal-sized EC-to-BT pieces.

Results Table 2 shows the results of these ablations. Evaluation is in terms of BLEU on the test set, and the Δ column reports the percent difference from the best value for a language in Table 1. We find significant reduction in translation quality with the CLIP image encoder and inconsistent performance for both an initial BT phase and interleaved training, with performance dropping for German but slightly increasing for Sinhala when compared to T2I-EC (as seen in the Δ column).

Manipulation	Lang	en→X	X→en	mean	Δ
CLIP-img	de	18.52	25.93	22.23	-1%
CLIP-img	si	1.05	4.18	2.61	-14%
Init BT	de	18.20	25.39	21.80	-3%
Init BT	si	1.24	4.84	3.04	+0.6%
Interleave	de	18.29	25.69	21.99	-2%
Interleave	si	1.25	4.84	3.05	+1%

Table 2: Results from several training pipeline manipulations. BLEU scores reported; Δ is the percentage difference from the corresponding mean value in T2I-EC in Figure 1.

6 Discussion

We have demonstrated that (at least one variant of) EC fine-tuning provides improvement on unsupervised translation over a standard backtranslation baseline. The gains are especially pronounced for the low-resource language Nepali, which is ideal since under-resourced languages constitute the expected use case for unsupervised translation techniques. Furthermore, since the hyperparameters for the EC-FT portion of our pipeline were mostly determined empirically, our approach may be *under-optimized*, meaning future work may yield further improvement using the same technique.

I2I-EC However, it is also clear that our formulation and implementation of "standard" EC (I2I-EC) does not improve upon the baseline, and even degrades performance in many cases. Our interpretation of this behavior is linked to our motivation for formulating T2I-EC in the first place.

As mentioned in Section 2.1, the image representations used in the sender's cross-attention, in the image-to-image setup, are not guaranteed to be at all similar to the representations that the receiver learns to encode. Because we seek to fine-tune for a standard *seq2seq* task (translation), it is desirable that the sender (mBART decoder) be trained to use the same or similar representations to those produced by the receiver (mBART encoder). Thus, we hypothesize that the null and negative effects of I2I-EC may be due to this mismatch between the representations the sender is trained to use, and those that the receiver is trained to produce.

However, we do **not** believe we have shown that I2I-EC will not be useful under slightly different formulations. In particular, the image representations may be able to be constrained to be similar to those of the receiver, either during EC or during the initial supervision phase. This could be accomplished using an auxillary distance loss, or by normalizing the mean and variance of the representations in both places.

EC Fine-Tuning Lastly, we view EC fine-tuning as a broader framework in which we have tested two distinct formulations (Steinert-Threlkeld et al., 2022). We will assume that the invariant element of EC is a model's use of discrete, natural-language generations as input to a second model, which must use them to accomplish some task.

Given this definition, there are several choice points for applying EC-FT. The parameter we explicitly explore in our experiments is whether the input to the sender is *image-based* or *text-based*. In both of our formulations, the receiver is trained by a contrastive image-choice loss. Another parameter for future work concerns whether this loss applies to images or texts. The receiver could be trained to choose the correct sentence out of a set of distractors via the similarity of the sentence embeddings.

A third parameter is whether the receiver is trained by a *contrastive* loss or a *generative* one (i.e. exactly reproducing a target sequence, as in *seq2seq* training).[8] In fact, an EC parameterization with text input, text output, and generative loss has already been formulated elsewhere, though it is not referred to as such. Niu et al. (2019) design a formulation of backtranslation, in which the artificial intermediate text is generated with straight-through

[8] Known as "reference game" versus "reconstruction game" in the EC literature (Lazaridou and Baroni, 2020).

Gumbel Softmax, instead of generated separately first. Future work will explore using this method with pre-trained models, i.e. in an EC-FT context.

These and other parameter choices leave extensive room for exciting future work with EC-FT as a general framework, both for UNMT and beyond.

7 Related Work

UNMT Unsupervised NMT uses only monolingual texts in each language of interest. Lample et al. (2018c) describe three principles for successful UNMT systems: 1. *initialization*, the initial model must leverage aligned representations between languages; 2. *language modeling*, there should be a strong "data driven prior" over the text patterns of each language; and 3. *backtranslation* which turns the unsupervised problem into a noisily-supervised one, through the use of semi-synthetic translations.

Significant progress has been made in improving each of these aspects of UNMT. Pre-trained multilingual language models (Lample and Conneau, 2019; Conneau et al., 2020; Liu et al., 2020; Song et al., 2019) have vastly improved the tractability of principles 1 and 2, largely replacing initialization techniques using inferred bilingual dictionaries (e.g. Lample et al., 2018b).

For the third principle, *iterative backtranslation* is widely used (Sennrich et al., 2016a; He et al., 2016a; Lample et al., 2018a; Haddow et al., 2022). On this approach, synthetic data is generated "on the fly", during training. The model is updated before each new batch of synthetic text is generated, leading to simultaneous incremental improvement in generated data quality and model quality.

In this work, we adhere to all three principles, but add EC as a training signal. It has been noted that UNMT baselines still perform relatively poorly for low-resource languages (Guzmán et al., 2019). We improve upon low-resource UNMT pipelines by leveraging goal-directed, multimodal fine-tuning via emergent communication.

EC and NLP A few other papers combine EC and NMT specifically. Lee et al. (2018) use EC and image captioning to build UNMT models, showing that EC promotes better translation than the multimodal alignment technique of Nakayama and Nishida (2017). Our approach differs in several important respects: we initialize our EC environment with *pre-trained language models*; we use both EC and backtranslation; and we do not simultaneously train on the EC objective and image captioning objective. Moreover, because we use one multilingual model, our caption grounding only uses one language, instead of all languages. Our results show that EC promotes unsupervised translation in the context of advanced methods that combine pre-training with backtranslation.

Li et al. (2020b) use emergent communication as a pre-training step for NMT systems. They have agents play an EC game, and then use those parameters to initialize an NMT system. They find that (together with adapters and weight-distance regularization) EC pre-training improves in BLEU over a standard NMT baseline, with especially large gains coming in the few-shot setting. While this shows that EC can provide a good initialization for a recurrent NMT system, our present work shows that EC can provide a good fine-tuning signal for a pre-trained multilingual language model. We also note two differences with respect to both works: (i) they use recurrent networks, whereas we start from a pre-trained transformer, and (ii) they use separate models for each language, whereas we use one multilingual model.

Lee et al. (2019) cast translation as a communication game with a third pivot language as the latent space in order to study (i) language drift from a pre-trained supervised MT model and (ii) using visual grounding (via gold image captions) plus language modeling to counter such drift. This approach thus does use EC with a pre-trained model, but it is a small model trained on the target task (translation). Our approach encourages using EC in conjunction with large-scale pre-trained language models which are intended to be general-purpose.

Finally, Lazaridou et al. (2020) study various ways of combining EC with a standard vision-language task, namely image captioning. They identify several forms of language drift and explore ways of incorporating auxillary losses. This work heavily inspires our own, since many of their settings correspond to using a pre-trained image-caption system. Our focus, however, has been on using EC to fine-tune large-scale pre-trained models on a language-only task, which introduces its own challenges and has its own benefits.

Multimodal pre-training Recently, efforts in multimodal pre-training are surging, especially in vision-language (V-L) pre-training (Du et al., 2022). Most of the works create joint V-L representations through a fusion encoder (Li et al., 2020a, 2019; Tan and Bansal, 2019), where the fused represen-

tation is the joint representation of image and text, as learned by a single encoder. Other recent works such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) attempt to use different encoders for images and text to make the framework more efficient. While V-L pre-training models image and text data jointly (Du et al., 2022; Wang et al., 2021), we start with an existing pre-trained language model and further train it through the communication process in an image referential game. Although we expect the alignment between image and text to arise through this process, we view the visual modality as an additional signal to ground the multilingual communication process.

We also note that most previous work on V-L pre-training is evaluated solely on vision or V-L tasks (Li et al., 2019; Radford et al., 2021; Jia et al., 2021). The advantage of this joint pre-training for language-only tasks remains unclear (Yun et al., 2021; Pezzelle et al., 2021). In this paper, we focus on a language-only task (UNMT) to evaluate whether visual grounding can improve such tasks.

Finally, we note that EC-FT is more general than typical approaches to multimodal pre-training. While the image-based task we employ here works by promoting multimodal alignment, the range of possible tasks that can be used in EC-FT is huge, from directing other agents (Mordatch and Abbeel, 2018) to controlling a robot (Das et al., 2019) to playing games and reasoning about social dilemmas (Jaques et al., 2019). This wide range of tasks can incorporate many dimensions of communication that should be beneficial for NLP systems—e.g. other agents with their own private goals, social context, embodied control—that are not easily captured by multimodal pre-training (Bender and Koller, 2020; Bisk et al., 2020). In terms of Bisk et al. (2020)'s *world scopes* mentioned in the introduction, multimodal pre-training corresponds to world scope 3 (perception); EC-FT has the ability to move us much closer towards the final scopes 4 (embodiment and action) and 5 (the social world).

Multimodal Fine-tuning A related body of work focuses on adapting pre-trained language-only models for use in multi-modal tasks. For example, Tsimpoukelli et al. (2021) show that using a frozen language model and adapting a visual encoder to produce embeddings aligned with the LM's can be useful for few-shot learning in multimodal tasks like visual question answering. Liang et al. (2022) make this approach more modular by additionally freezing the visual encoder and learning separate prompt vectors. In the EC-FT context, these works suffer some of the same limitations in world scope, but could provide very useful methods for the environment-to-sender adapter step discussed in Section 2.1.

8 Conclusion

We have shown that Emergent Communication can be used as a fine-tuning signal for a large pre-trained multilingual model; this grounding in a goal-oriented multimodal task yields improvements over an unsupervised NMT baseline in all four languages studied. There is likely room to further improve upon the specific EC variants we propose here, since we believe the EC process is under-optimized for hyperparameters. We have further noted that the framework we propose leaves extensive room for further experimentation, since there are many choice points of the general EC setup that we have not yet tested, and may be promising avenues for future improvement. The general EC-FT framework may also be applied to other tasks beyond UNMT in future work.

Author Contribution Statement

Following a practice in several other fields, we here list author contributions according to the Contributor Role Taxonomy (CRediT; Allen et al., 2019). **C.M. Downey:** Conceptualization, Methodology, Software, Validation, Investigation, Writing - original draft, Writing - review and editing, Visualization. **Xuhui Zhou:** Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing - review and editing. **Zeyu Liu:** Conceptualization, Methodology, Software, Validation, Investigation, Data curation, Writing - review and editing. **Shane Steinert-Threkeld:** Conceptualization, Methodology, Resources, Writing - original draft, Writing - review and editing, Supervision, Project administration, Funding acquisition.

Limitations

One limitation of our work concerns analysis. Much remains to be learned about the mechanisms by which EC can help translation. By evaluating the model more comprehensively, we could gain insight into whether and how the grounding helps task performance. Based on such analysis, a better version of the pipeline could be developed.

We observed significant variability across random seeds in our EC training; methods for stabilizing this variability could ensure the reliability of EC as a fine-tuning process for models.

Finally, we investigated only four non-English languages, two 'high-resource' and two 'low-resource'. It would be valuable to explore a wider range of typologically diverse languages to validate that these methods apply across the board and, if not, to understand what language factors drive success.

Ethics Statement

This work on unsupervised translation should have a positive impact on many under-served language communities by extending the reach of a core language technology (translation) to languages which lack the extensive parallel data required for supervised translation systems.

That being said, there are ethical risks with the present approach. The pre-training of mBART depends on the CommonCrawl dataset, so there might be some offensive language and even identity leakage due to CommonCrawl's preprocessing pipeline. It is possible that the model will generate toxic and biased utterances in our experiments. We didn't evaluate the toxicity of our generation. Our intuition is that the caption grounding will bias the model towards descriptive captions and thus suppress the toxic generation.

Acknowledgments

We thank Emily M Bender, Emmanuel Chemla, Chris Potts, Tania Rojas-Esponda, and the anonymous reviewers of and audience at the ICLR 2022 Emergent Communication workshop for helpful discussion. This work was partially supported by funding from the University of Washington Royalty Research Fund (RRF), grant number A167354 "Learning to translate by learning to communicate".

References

Liz Allen, Alison O'Connell, and Veronique Kiermer. 2019. How can we ensure visibility and diversity in research contributions? How the Contributor Role Taxonomy (CRediT) is helping the shift from authorship to contributorship. *Learned Publishing*, 32(1):71–74.

Jacob Andreas. 2019. Measuring compositionality in representation learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *International Conference of Learning Representations*. _eprint: 1710.11041.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference of Learning Representations (ICLR)*, pages 1–15. _eprint: 1409.0473.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online. Association for Computational Linguistics.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models Are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and Generalization In Emergent

Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4427–4442, Stroudsburg, PA, USA. Association for Computational Linguistics. _eprint: 2004.09124.

Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2019a. Anti-efficient encoding in emergent communication. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. _eprint: 1905.12561v4.

Rahma Chaabouni, Eugene Kharitonov, Alessandro Lazaric, Emmanuel Dupoux, and Marco Baroni. 2019b. Word-order Biases in Deep-agent Emergent Communication. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5166–5175, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. 2019. TarMAC: Targeted Multi-Agent Communication. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1538–1546. PMLR.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference of Learning Representations (ICLR)*.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. *CoRR*, abs/2202.10936.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Nur Geffen Lan, Emmanuel Chemla, and Shane Steinert-Threlkeld. 2020. On the Spontaneous Emergence of Discrete and Compositional Signals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4794–4800, Stroudsburg, PA, USA. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. Technical Report arXiv:2109.00486, arXiv. ArXiv:2109.00486 [cs] type: article.

Stevan Harnad. 1990. The Symbol Grounding Problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346.

Serhii Havrylov and Ivan Titov. 2017. Emergence of Language with Multi-agent Games: Learning to Communicate with Sequences of Symbols. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*. _eprint: 1705.11192.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016a. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Eric Jang, Shixiang Gu, and Ben Poole. 2017. Categorical Reparameterization with Gumbel-Softmax. In *Proceedings of the International Conference on Learning Representations*.

Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, Dj Strouse, Joel Z. Leibo, and Nando De Freitas. 2019. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *Proceedings of the 36th*

International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pages 3040–3049. PMLR.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Stroudsburg, PA, USA. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual Language Model Pretraining. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. _eprint: 1901.07291.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference of Learning Representations (ICLR)*.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018b. Word translation without parallel data. In *International Conference on Learning Representations*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018c. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Angeliki Lazaridou and Marco Baroni. 2020. Emergent Multi-Agent Communication in the Deep Learning Era. pages 1–24. _eprint: 2006.02419.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Douwe Kiela. 2019. Countering language drift via visual grounding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4385–4395, Hong Kong, China. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. 2018. Emergent translation in multi-agent communication. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pages 1–18. _eprint: 1710.06922.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Stroudsburg, PA, USA. Association for Computational Linguistics. _eprint: 1910.13461.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Yaoyiran Li, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2020b. Emergent communication pretraining for few-shot machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4716–4731, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sheng Liang, Mengjie Zhao, and Hinrich Schütze. 2022. Modular and Parameter-Efficient Multimodal Fusion with Prompting. *arXiv:2203.08055 [cs]*. ArXiv: 2203.08055.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. *CoRR*, abs/1405.0312. ArXiv: 1405.0312.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer*

Vision – ECCV 2018, volume 11219, pages 353–369. Springer International Publishing, Cham.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Ryan Lowe, Abhinav Gupta, Jakob Foerster, Douwe Kiela, and Joelle Pineau. 2020. On the interaction between supervision and self-play in emergent communication. In *International Conference on Learning Representations*.

Yuchen Lu, Soumye Singhal, Florian Strub, Aaron Courville, and Olivier Pietquin. 2020. Countering Language Drift with Seeded Iterated Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6437–6447. PMLR.

Chris J Maddison, Andriy Mnih, Yee Whye Teh, United Kingdom, and United Kingdom. 2017. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference of Learning Representations (ICLR)*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

William Merrill, Yoav Goldberg, Roy Schwartz, and Noah A. Smith. 2021. Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand? *Transactions of the Association for Computational Linguistics*, 9:1047–1060.

Igor Mordatch and P. Abbeel. 2018. Emergence of Grounded Compositional Language in Multi-Agent Populations. In *AAAI*.

Jesse Mu and Noah D. Goodman. 2021. Emergent Communication of Generalizations. In *Proceedings of Neural Information Processing Systems (NeurIPS)*.

James Mullenbach, Yada Pruksachatkun, Sean Adler, Jennifer Seale, Jordan Swartz, Greg McKelvey, Hui Dai, Yi Yang, and David Sontag. 2021. CLIP: A dataset for extracting action items for physicians from hospital discharge notes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1365–1378, Online. Association for Computational Linguistics.

Hideki Nakayama and Noriki Nishida. 2017. Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *Machine Translation*, 31(1-2):49–64. Publisher: Springer Netherlands _eprint: 1611.04503.

Xing Niu, Weijia Xu, and Marine Carpuat. 2019. Bidirectional differentiable input reconstruction for low-resource neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 442–448, Minneapolis, Minnesota. Association for Computational Linguistics.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word Representation Learning in Multimodal Pre-Trained Transformers: An Intrinsic Evaluation. *Transactions of the Association for Computational Linguistics*, 9:1563–1579.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Brian Skyrms. 2010. *Signals: Evolution, Learning, and Information*. Oxford University Press.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5926–5936. PMLR.

Shane Steinert-Threlkeld. 2020. Toward the Emergence of Nontrivial Compositionality. *Philosophy of Science*, 87(5):897–909. Publisher: The University of Chicago Press.

Shane Steinert-Threlkeld, Leo Z. Liu, Xuhui Zhou, and C.M. Downey. 2022. Emergent communication fine-tuning (EC-FT) for pretrained language models. In *Proceedings of the 5th Annual Workshop on Emergent Communication, ICLR*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*. _eprint: 1409.3215.

Mariarosaria Taddeo and Luciano Floridi. 2005. Solving the Symbol Grounding Problem: A Critical Review of Fifteen Years of Research. *Journal of Experimental & Theoretical Artificial Intelligence*, 17(4):419–445. Publisher: Taylor & Francis.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Maria Tsimpoukelli, Jacob Menick, and Serkan Cabi. 2021. Multimodal Few-Shot Learning with Frozen Language Models. In *Neural Information Processing Systems*.

Kyle Wagner, James A. Reggia, Juan Uriagereka, and Gerald Wilkinson. 2003. Progress in the Simulation of Emergent Communication and Language. *Adapt. Behav.*, 11(1):37–69.

Jianfeng Wang, Xiaowei Hu, Zhe Gan, Zhengyuan Yang, Xiyang Dai, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2021. UFO: A unified transformer for vision-language representation learning. *CoRR*, abs/2111.10023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, \Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. _eprint: 1609.08144.

Tian Yun, Chen Sun, and Ellie Pavlick. 2021. Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4357–4366, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Main Experiments Training Details

We here include more details about the training protocol for the results reported in Section 4. Our codebase is built upon the mBART code from `huggingface` (Wolf et al., 2019) and PyTorch (Paszke et al., 2019). We use one NVIDIA RTX 8000 GPU for each experiment. Backtranslation is the most expensive part of the entire training pipeline. It takes around 24-28 hours to finish, depending on the languages. The combined training

time for caption grounding and emergent communication is within 1 hour.

Baseline As discussed in section 3, our UNMT baseline is established by starting with mBART and performing 8192 steps of iterative backtranslation for each translation pair. We use a batch size of 32 and a maximum generated sequence length of 64. See more hyperparameter choices in Table 3.

I2I-EC For our I2I-EC fine-tuned model, training consists of the following pipeline

1. 2048 steps of backtranslation
2. 2048 steps of supervised captioning training (English-only)
3. 2048 steps of EC fine-tuning
4. 6144 steps of backtranslation

Backtranslation uses the same exact hyperparameters as in the baseline, but with training split between the first 2048 and last 6144 steps (Table 3).

Supervised caption training is described in Section 2.1. We have 8 choices for the image selection task (7 distractors and 1 correct choice). As part of Sender agent, we use a one-layer autoregressive transformer to serialize (or, "unroll") a single ResNet image representation to a sequence of vectors to imitate the sequential data mBART observes during its pre-training. The unrolled sequence is used in the sender's cross-attention, and the sender is trained to generate the gold-standard caption.

Also during the supervised captioning stage, the receiver takes in the gold-standard caption, and a one-layer RNN is used to aggregate its final hidden states and choose the correct image. The image selection (cross-entropy) loss is scaled with λ, before being added to the caption-generation loss. Full hyperparameter choices are detailed in Table 4.

I2I-EC fine-tuning is also described in Section 2.1. Different from caption grounding, we have a total of 16 image choices instead of 8. The adapter unrolls the ResNet image representation to a length of 32. The emergent generation is language-constrained as described in Section 3 with a threshold. A repetition penalty is applied to the generations, and they are constrained to not repeat any 4-grams or longer. KL-regularization with a separate mBART instance fine-tuned on causal language modeling is applied with a λ parameter. Full hyperparameter choices are detailed in Table 5.

T2I-EC For our T2I-EC fine-tuned model, training is performed slightly differently for empirical reasons

1. 2048 steps of supervised captioning training (English-only)
2. 2048 steps of EC fine-tuning
3. 8192 steps of backtranslation

T2I-EC hyperparameters are very similar to I2I-EC. See full parameters in Tables 3, 4, and 5.

Auxiliary CLM To have a language model for use in KL regularization (see equation (2)), we fine-tuned just the mBART decoder on the same common crawl data used for its pretraining in all of the languages of interest. We trained for 100000 steps, batch size 32, sequence length 96, and learning rate of 6×10^{-6}. This model was then frozen during EC training and only used to compute the KL divergence which was used in updating the sender's parameters.

B Manipulations Training Details

All manipulations are performed on the main T2I-EC process. Interleaved training uses versions of the the learning rate schedules used for the main experiments shortened by a factor of 4.

C Full results

In Table 6, we include full results for our main experiment (summarized in Table 1). Although we found the EC process to help with machine translation, it also leads to instability in model training. We a systematic study of this variation to future work.

In Table 7 we show experiments with a more modern choice of image encoder — CLIP-Large (Mullenbach et al., 2021). We find that the CLIP-Large encoder under-performs ResNet.

The full results from our manipulation experiments (Section 5) are found in Table 8.

Name	Values
optimizer	`Adam(betas=(0.9, 0.999))` (default in PyTorch)
LR scheduler	`constant_w_warmup`
`grad_clip`	1.0
`batch_size`	32
`evaluate_bleu_every`	256
|validation_set|	4096
#beams	5
first `#vocab_constrained_steps`	2048
threshold (after `#vocab_constrained_steps`)	0.99
#warmup_steps	$\frac{1}{4}\cdot$ #steps

(a) Backtranslation shared parameters

(b) Baseline

Name	Values
Learning rate	2.0e-5
#steps	8192
first_threshold	0.90

(c) I2I-EC (Initial BT)

Name	Values
Learning rate	1.0e-5
#steps	2048
first_threshold	0.96

(d) I2I-EC (Secondary BT)

Name	Values
Learning rate	1.0e-5
#steps	6144

(e) T2I-EC

Name	Values
Learning rate	1.0e-5
#steps	8192
first_threshold	0.96

Table 3: Hyper-parameters for backtranslation.

Name	Values
optimizer	`Adam(betas=(0.9, 0.999))` (default in PyTorch)
#steps	2048
learning rate	4.0e-5
LR scheduler	`linear_w_warmup`
#warm-up steps	0
batch_size	16
#distractors	7
Reshaper (Sender & Receiver)	linear projection
Dropout (anywhere)	0.0
Image Unroll	one (auto-regressive) transformer layer
Image Unroll length	32
Receiver aggregation	RNN
Sender	no freezing
Receiver	no freezing
beam_width	1 (Greedy)
temperature	1.0
gumble_softmax sample	one-hot
`repetition_penalty`	1.0
`max_seq_length`	32

(a) Captioning shared parameters

(b) I2I-EC

Name	Values
Image selection loss λ	4.0
grad_clip	1.0

(c) T2I-EC

Name	Values
Image selection loss λ	8.0
grad_clip	0.5

Table 4: Hyper-parameters for caption grounding part of emergent communication.

Name	Values
optimizer	`Adam(betas=(0.9, 0.999))` (default in PyTorch)
#steps	2048
LR scheduler	`linear_w_warmup`
#warm-up steps	0
batch_size	12
#distractors	15
Reshaper (Sender & Receiver)	linear projection
Dropout (anywhere)	0.0
Image Unroll	one (auto-regressive) transformer layer
Image Unroll length	32
Receiver aggregation	RNN
Sender	no freezing
Receiver	no freezing
beam_width	1 (Greedy)
temperature	1.0
gumble_softmax sample	one-hot
`vocab_constraint_threshold`	0.99
`repetition_penalty`	1.0
`max_seq_length`	32

(a) Emergent communication shared parameters

(b) I2I-EC

Name	Values
Language modeling loss λ	0.125
Learning rate	6.0e-6
grad_clip	1.0

(c) T2I-EC. *: length of text string in place of series of "pseudo-images" from image unroller

Name	Values
Language modeling loss λ	0.0625
Learning rate	1.0e-6
grad_clip	0.5
`max_text_seq_length`*	128

Table 5: Hyper-parameters for emergent communication.

Model	Language	Seed	BLEU			COMET		
			en→X	X→en	mean	en→X	X→en	mean
baseline (mBART + BT)	zh	1	17.21	11.35	14.28	-0.04	0.14	0.05
		2	18.38	11.39	14.89	0.02	0.14	0.08
		3	18.45	11.36	14.90	0.03	0.15	0.09
	de	1	18.66	25.83	22.24	0.18	0.39	0.29
		2	19.06	25.73	22.39	0.20	0.38	0.29
		3	18.79	25.88	22.33	0.22	0.40	0.31
	ne	1	1.94	4.74	3.34	-0.19	-0.36	-0.27
		2	1.84	4.94	3.39	-0.20	-0.34	-0.27
		3	2.14	5.07	3.60	-0.24	-0.34	-0.29
	si	1	1.29	4.53	2.91	-0.29	-0.31	-0.30
		2	1.18	4.73	2.95	-0.18	-0.28	-0.23
		3	1.21	4.35	2.78	-0.20	-0.32	-0.26
I2I-EC	zh	1	17.31	10.96	14.13	-0.03	0.12	0.05
		2	17.03	11.24	14.14	0.00	0.15	0.07
		3	18.72	11.88	15.30	0.04	0.17	0.10
	de	1	18.22	25.41	21.81	0.18	0.39	0.29
		2	18.26	25.60	21.93	0.18	0.39	0.29
		3	18.06	25.28	21.67	0.20	0.40	0.30
	ne	1	1.24	5.13	3.19	-0.25	-0.31	-0.28
		2	1.22	5.30	3.26	-0.25	-0.36	-0.31
		3	1.51	5.34	3.43	-0.24	-0.33	-0.29
	si	1	0.01	0.08	0.04	-1.63	-1.05	-1.34
		2	0.00	0.02	0.01	-1.31	-1.28	-1.30
		3	0.01	0.05	0.03	-1.40	-1.15	-1.28
T2I-EC	zh	1	19.25	11.91	15.58	0.06	0.18	0.12
		2	0.09	0.11	0.10	-1.75	-1.60	-1.68
		3	18.60	12.27	15.43	0.05	0.18	0.11
	de	1	17.91	25.72	21.81	0.18	0.38	0.28
		2	18.64	26.20	22.42	0.19	0.41	0.30
		3	18.56	25.82	22.19	0.19	0.39	0.29
	ne	1	0.06	0.03	0.04	-1.27	-1.14	-1.20
		2	0.02	0.11	0.07	-1.33	-1.06	-1.20
		3	2.36	5.92	4.14	-0.20	-0.27	-0.24
	si	1	1.10	4.33	2.72	-0.25	-0.29	-0.27
		2	0.01	0.19	0.10	-1.42	-1.12	-1.27
		3	1.28	4.76	3.02	-0.18	-0.27	-0.22

Table 6: Full results of our main experiment with ResNet image representation.

Model	Language	Seed	BLEU			COMET		
			en→X	X→en	mean	en→X	X→en	mean
baseline (mBART + BT)	zh	1	17.21	11.35	14.28	-0.04	0.14	0.05
		2	18.38	11.39	14.89	0.02	0.14	0.08
		3	18.45	11.36	14.90	0.03	0.15	0.09
	de	1	18.66	25.83	22.24	0.18	0.39	0.29
		2	19.06	25.73	22.39	0.20	0.38	0.29
		3	18.79	25.88	22.33	0.22	0.40	0.31
	ne	1	1.94	4.74	3.34	-0.19	-0.36	-0.27
		2	1.84	4.94	3.39	-0.20	-0.34	-0.27
		3	2.14	5.07	3.60	-0.24	-0.34	-0.29
	si	1	1.29	4.53	2.91	-0.29	-0.31	-0.30
		2	1.18	4.73	2.95	-0.18	-0.28	-0.23
		3	1.21	4.35	2.78	-0.20	-0.32	-0.26
I2I-EC	zh	1	16.66	10.94	13.80	-0.07	0.13	0.03
		2	17.46	10.87	14.16	-0.01	0.13	0.06
		3	18.84	11.64	15.24	0.03	0.16	0.10
	de	1	18.64	26.17	22.40	0.22	0.40	0.31
		2	17.98	25.20	21.59	0.20	0.38	0.29
		3	18.09	25.35	21.72	0.20	0.40	0.30
	ne	1	1.02	4.68	2.85	-0.41	-0.38	-0.39
		2	1.87	5.19	3.53	-0.26	-0.33	-0.29
		3	1.79	5.29	3.54	-0.20	-0.34	-0.27
	si	1	0.30	1.64	0.97	-1.14	-0.59	-0.87
		2	0.16	0.55	0.36	-0.88	-0.88	-0.88
		3	0.76	4.88	2.82	-0.37	-0.29	-0.33
T2I-EC	zh	1	0.04	0.09	0.07	-1.69	-1.43	-1.56
		2	17.77	12.02	14.90	0.00	0.18	0.09
		3	17.24	11.23	14.24	-0.03	0.13	0.05
	de	1	10.45	14.14	12.29	-0.42	-0.30	-0.36
		2	18.52	25.93	22.23	0.20	0.40	0.30
		3	18.26	25.61	21.94	0.19	0.38	0.28
	ne	1	0.75	2.49	1.62	-0.85	-0.58	-0.71
		2	0.09	0.07	0.08	-1.37	-1.18	-1.28
		3	0.02	0.04	0.03	-1.35	-1.17	-1.26
	si	1	0.02	0.15	0.09	-2.00	-1.45	-1.72
		2	0.04	0.19	0.12	-2.02	-1.32	-1.67
		3	1.05	4.18	2.61	-0.33	-0.28	-0.30

Table 7: Full results of our main experiment with CLIP-Large image representation.

Manipulation	Language	Seed	BLEU			COMET		
			en→X	X→en	mean	en→X	X→en	mean
CLIP-img	de	1	10.45	14.14	12.29	-0.42	-0.30	-0.36
		2	18.52	25.93	22.23	0.20	0.40	0.30
		3	18.26	25.61	21.94	0.19	0.38	0.28
	si	1	0.02	0.15	0.09	-2.00	-1.45	-1.72
		2	0.04	0.19	0.12	-2.02	-1.32	-1.67
		3	1.05	4.18	2.61	-0.33	-0.28	-0.30
Init BT	de	1	18.49	25.87	22.18	0.19	0.40	0.30
		2	17.28	24.89	21.08	0.12	0.32	0.22
		3	18.20	25.39	21.80	0.22	0.40	0.31
	si	1	0.94	4.56	2.75	-0.43	-0.27	-0.35
		2	1.24	4.84	3.04	-0.28	-0.25	-0.27
		3	0.09	0.62	0.35	-1.24	-0.84	-1.04
Interleave	de	1	18.23	25.56	21.90	0.15	0.39	0.27
		2	18.29	25.69	21.99	0.18	0.38	0.28
		3	17.93	25.81	21.87	0.16	0.39	0.27
	si	1	0.01	0.02	0.02	-1.57	-1.34	-1.46
		2	1.25	4.84	3.05	-0.34	-0.25	-0.30
		3	1.04	4.37	2.70	-0.46	-0.28	-0.37

Table 8: Results from several T2I-EC training pipeline manipulations.