

Choosing What to Mask: More Informed Masking for Multimodal Machine Translation

Júlia Sato*, Helena Caseli*, Lucia Specia†

*Federal University of São Carlos (UFSCar), São Carlos, Brazil

†Imperial College London, London, United Kingdom

juliasato@estudante.ufscar.br

helenacaseli@ufscar.br

l.specia@ic.ac.uk

Abstract

Pre-trained language models have achieved remarkable results on several NLP tasks. Most of them adopt masked language modeling to learn representations by randomly masking tokens and predicting them based on their context. However, this random selection of tokens to be masked is inefficient to learn some language patterns as it may not consider linguistic information that can be helpful for many NLP tasks, such as multimodal machine translation (MMT). Hence, we propose three novel masking strategies for cross-lingual visual pre-training – more informed visual masking, more informed textual masking, and more informed visual and textual masking – each one focusing on learning different linguistic patterns. We apply them to Vision Translation Language Modelling for video subtitles (Sato et al., 2022) and conduct extensive experiments on the Portuguese-English MMT task. The results show that our masking approaches yield significant improvements over the original random masking strategy for downstream MMT performance. Our models outperform the MMT baseline and we achieve state-of-the-art accuracy (52.70 in terms of BLEU score) on the How2 dataset, indicating that more informed masking helps in acquiring an understanding of specific language structures and has great potential for language understanding¹.

1 Introduction

Pre-trained language models have achieved remarkable results on several Natural Language Processing (NLP) tasks (Devlin et al., 2019; Liu et al., 2019; Baevski et al., 2019; Yang et al., 2019; Joshi et al., 2020; Clark et al., 2020; Lan et al., 2020; Zhuang et al., 2021). One of these tasks is multimodal machine translation (MMT), which has attracted considerable attention from both Computer

Vision and NLP communities as it not only considers text information but also uses other modal information – mostly visual information – to improve translation outputs (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Recent advances in this field have achieved significant success and highlighted the efficiency of both multimodal and multilingual pre-training for MMT (Caglayan et al., 2021; Sato et al., 2022).

Nonetheless, most pre-trained models follow BERT’s pre-training paradigm (Devlin et al., 2019) and adopt masked language modeling (MLM) and its variants to learn representations by masking tokens and making predictions based on their context. The conventional MLM relies on randomly selecting tokens to be masked and therefore may not consider linguistic information that can be helpful for some NLP tasks, such as MMT.

In this paper, we address this problem through a systematic study of new masking approaches for cross-lingual visual pre-training. We propose *more informed* masking strategies to learn particular language patterns for downstream multimodal machine translation performance. These strategies consist of selectively masking linguistic and visual tokens instead of randomly masking them, focusing on situations that can be favored by a better understanding of specific visual or textual information.

For instance, since most pre-trained language models are based on English, they fail to understand some linguistic patterns that are common in many other languages, such as the grammatical gender of words. The English language treats the grammatical gender of words differently from languages such as French, Spanish, Portuguese, or Italian. While some languages have different words with the same meaning that are found in the feminine and masculine forms, this does not happen in the English language. For example, considering the English-Portuguese translation, the pronoun

¹The source codes have been released at <https://github.com/LALIC-UFSCar/more-informed-masking>

“they” can be translated to “elas” (feminine) or “eles” (masculine). Another example is the adjective “beautiful”, which can be translated to “bonita” (feminine) or “bonito” (masculine) depending on who or what it is referring to.

In this context, we propose three selective masking strategies – more informed visual masking, more informed textual masking, and more informed visual and textual masking – each one focusing on masking specific linguistic and visual tokens that can contribute to better understanding some of these different linguistic patterns. We apply them to Vision Translation Language Modelling for video subtitles (Sato et al., 2022) and run an extensive set of experiments on the Portuguese-English MMT task.

We find that predicting particular masked elements can be a powerful objective for cross-lingual visual pre-training as the pre-trained model can acquire a better understanding of specific language structures. Experimental results show that our masking approaches yield significant improvements over the original random masking strategy for downstream MMT performance. Our models outperform the MMT baseline and achieve state-of-the-art accuracy (52.70 in terms of BLEU score) on the How2 dataset (Sanabria et al., 2018), indicating that more informed masking helps in capturing domain-specific language patterns and has great potential for language understanding.

2 Method

In this section, we present the detailed implementation of three masking strategies: more informed visual masking (Section 2.2.1), more informed textual masking (Section 2.2.2), and more informed visual and textual masking (Section 2.2.3), as well as the VTLM for video subtitles pre-training objective in Section 2.1.

2.1 Visual translation language modelling for video subtitles

The VTLM objective (Caglayan et al., 2021) joins the translation language modelling (TLM) (Conneau and Lample, 2019), which employs the masked language modelling objective, with masked region classification (MRC) (Chen et al., 2020; Su et al., 2020) to generate cross-lingual and multimodal representations. VTLM defines the input x as the concatenation of m -length source language sentence $s_{1:m}^{(1)}$, n -length target language sentence

$s_{1:n}^{(2)}$, and $\{v_1, \dots, v_o\}$ corresponding image features:

$$x = [s_1^{(1)}, \dots, s_m^{(1)}, s_1^{(2)}, \dots, s_n^{(2)}, v_1, \dots, v_o]$$

The final model combines the TLM loss with the MRC loss according to the following equation:

$$\mathcal{L} = \frac{1}{|X|} \sum_{x \in \mathcal{X}} \log Pr(\{\hat{y}, \hat{v}\} | \tilde{x}; \theta)$$

where \tilde{x} is the masked input sequence, \hat{y} denotes the ground-truth targets for masked positions, \hat{v} represents the detection labels and θ denotes the model parameters.

VTLM for video subtitles (Sato et al., 2022) corresponds to VTLM adapted to the Brazilian Portuguese-English language pair and to more challenging circumstances regarding the image-text relationship. Its pre-training has visual and cross-lingual resources and performs MLM and MRC on a three-way parallel multimodal and multilingual corpus, How2 (Sanabria et al., 2018).

Masking. VTLM selects a random set of linguistic and visual input tokens for masking. The masking proportion is 15% and it is applied separately to visual and language flows. For textual masking, 80% of the 15% chosen tokens are replaced with the [MASK] token, 10% are replaced with random tokens from the vocabulary, and 10% are kept unchanged. And visual masking follows a similar approach: VTLM replaces its vector of projected features by the [MASK] token embedding, with 10% of the masking being equivalent to using region features randomly selected from all images in the batch, and the remaining 10% of the regions are left intact.

2.2 More informed masking strategies

Unlike the original approach, we do not randomly select tokens for masking. Instead, we focus on masking specific tokens in order to learn particular language patterns efficiently. Thus, we propose three new masking strategies that explore more informed ways of masking linguistic and visual tokens.

These approaches are based on the hypothesis that by performing more informed masking (e.g., masking tokens that reveal the grammatical gender of words) the model could come to a better understanding of these concepts, obtaining better performance in the translation of pronouns and words assigned as masculine, feminine, or neuter.

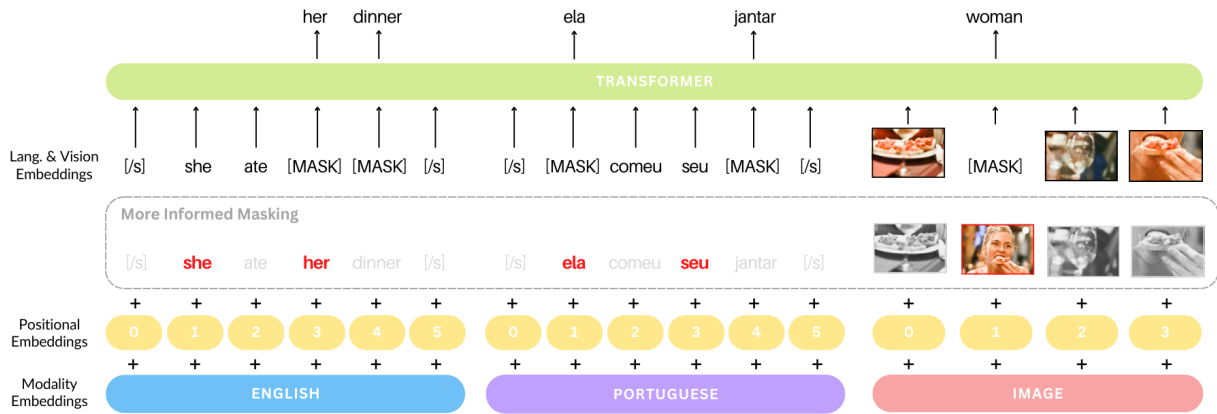


Figure 1: VTLM architecture, highlighting the *more informed visual and textual masking* strategy.

The overall architecture of the model is depicted in Figure 1.

2.2.1 More informed visual masking

This approach consists of changing the visual masking so that the initial selection of tokens for masking is no longer random, and a greater proportion of tokens related to elements categorized as *people* are selected for masking, such as objects in the image categorized as “man”, “woman”, “boy”, or “girl”. For convenience, we denote these tokens as T_{People} .

To accomplish this, we changed the visual masking stage to retrieve detection information necessary to perform the identification of class labels during training. Specifically, we used object features that were previously extracted using the Faster R-CNN model (Ren et al., 2015) pre-trained on the Open Images Dataset V4 (Kuznetsova et al., 2020) to retrieve the information needed to identify the categories of visual tokens during training.

At the beginning of the visual masking stage, we obtain the category index from the label map of the Open Images Dataset, as well as the variables containing the class predictions and confidence scores for each image from the batch. We then identify the index associated with each image and the position of each visual token in relation to the set of images from the batch. As a result, we are able to obtain the class label and confidence score for each token candidate to be masked and selectively choose the tokens that will be masked.

We apply this strategy to increase the proportion of T_{People} among masked tokens, with a percentage of 33.34%, 50.0%, and 66.67%. In all cases, the remaining candidate tokens for making do not have the same category as T_{People} and are randomly

chosen. We maintained the visual masking ratio: 15% of inputs are selected for masking, from which 80% are replaced with the [MASK] token, 10% are replaced with random tokens, and 10% are left intact.

2.2.2 More informed textual masking

Similar to the previous approach, this masking strategy aims to mask a greater amount of tokens that reveal the grammatical gender of words in a given sentence. Thus, the initial selection of tokens for masking was changed to no longer be random and to favor more pronouns – such as “he”/“she”, “him”/“her”, and “his”/“hers” – among the tokens that will be masked, maintaining the 15% textual masking ratio. For convenience, we denote these tokens as T_{Pronouns} .

As VTLM stores the input textual stream as integer-type *Tensors*, we changed the VTLM architecture to convert this numerical stream to words at the beginning of the textual masking stage and then ascertain each sentence from the batch to identify subject pronouns, object pronouns, and possessive adjectives and pronouns. After identifying these words, they are marked and associated with their original numerical form so that they can be identified later in the selection of tokens for masking. At this stage, T_{Pronouns} are identified and tokens are selectively chosen to be masked, with a higher proportion of T_{Pronouns} being masked.

We performed three experiments with the following percentages of T_{Pronouns} : 33.34%, 50.0%, and 66.67%. In all cases, the remaining masked tokens did not have the same category as T_{Pronouns} and were randomly chosen following the standard approach.

Model	T_{People}	Test		Valid	
		BLEU	METEOR	BLEU	METEOR
VTLM: random masking		51.80	78.04	52.44	78.25
	33.34%	52.70	79.63	53.25	79.83
	50.00%	51.92	79.10	52.51	79.41
VTLM: more informed visual masking	66.67%	51.65	78.64	52.26	79.09

Table 1: BLEU and METEOR scores for random masking VTLM (baseline) and more informed visual masking VTLM (our model) for the MMT task.

2.2.3 More informed visual and textual masking

The more informed visual and textual masking strategy is a combination of the two previous approaches, i.e., we mask a greater proportion of T_{People} tokens at the visual masking stage, as well as T_{Pronouns} tokens at the textual masking stage.

This approach aimed to analyze the model behavior when applying more informed visual masking and more informed textual masking simultaneously.

3 Experiments

Pre-training data. We use the How2 corpus (Sanabria et al., 2018) in all stages of experimentation. How2 is a multimodal and multilingual collection of approximately 80,000 instructional videos accompanied by English subtitles and around 300 hours of collected crowdsourced Portuguese translations. For pre-training, we used a set from the How2 corpus that contains 155k features and their corresponding text in English and Portuguese². We applied Moses tokenization³ and used byte pair encoding (Sennrich et al., 2016) to split words into subword units.

Pre-training. We followed Caglayan et al.’s (2021) work to conduct the experiments. We set the model dimension to 512, the feed-forward layer dimension to 2048, the number of layers to 6 and the number of attention heads to 8. We randomly initialize model parameters rather than using pre-trained LM checkpoints. We use Adam (Kingma and Ba, 2014) with the mini-batch size set to 32 and the learning rate set to 0.0001. We set the dropout (Srivastava et al., 2014) rate to 0.1 in all layers. The pre-training was conducted on a single NVIDIA GeForce GTX 1070 GPU for 1.5M steps, and best

²The dataset used in this work is publicly available under the Creative Commons BY-SA 4.0 License and BSD-2-Clause License.

³<https://github.com/moses-smt/mosesdecoder>

checkpoints were selected with respect to validation set accuracy.

Fine-tuning. The encoder and the decoder of Transformer-based (Vaswani et al., 2017) MMT models are initialized with weights from VTLM, and fine-tuned with a smaller learning rate. We use the same hyperparameters as the pre-training phase, but we follow Sato et al.’s (2022) work and decrease the batch size to 16 and the learning rate to 1e-5. For evaluation, we use the models with the lowest validation set perplexity to decode translations with beam size of 8.

Evaluation Metrics. We report the automatic evaluation using BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005). We also conduct qualitative analyzes to better show the effects of the proposed masking strategies.

4 Results

The trained models were evaluated on valid and test sets of How2 for the multimodal machine translation (MMT) task. We compare our models with the original VTLM for video subtitles model (Sato et al., 2022), which has the same architecture but uses the popular random masking strategy instead of ours.

4.1 More informed visual masking

Table 1 shows BLEU and METEOR scores across valid and test sets of How2. The results show that this new masking strategy affects the final performance of the model. For $T_{\text{People}} = 33.34\%$, our model achieved 52.70 BLEU and 79.63 METEOR on the test set and 53.25 BLEU and 79.83 METEOR on the valid set for the MMT task, outperforming the baseline by approximately 1 BLEU and 1.6 METEOR. When $T_{\text{People}} = 50.0\%$, our model also outperformed the baseline in terms of both BLEU and METEOR, but its performance was slightly inferior to the performance of the first experiment. Finally, when $T_{\text{People}} = 66.67\%$, the

	<p>Source: Então ele ou ela não carrega todo o peso do SCBA, na área do ombro ou região ao redor do pescoço.</p> <p>Reference: So he or she is not carrying all the weight of the SCBA, in the shoulder area, or region around the neck.</p> <p>Baseline: So it or it doesn't carry all the weight of the SCBA, in the shoulder area, or region around the neck.</p> <p>Our model: So he or she won't carry all the weight of the SCBA, in the shoulder area, or region around the neck.</p>
	<p>Source: Então, há algumas maneiras diferentes de levá-lo pra fora.</p> <p>Reference: So there's a couple of different ways to take him out.</p> <p>Baseline: So there's a couple of different ways to take it out.</p> <p>Our model: So there's a couple of different ways to get him out.</p>
	<p>Source: E nós vamos fazer isso em seu cabelo hoje.</p> <p>Reference: And we're going to be cornrowing that into her hair today.</p> <p>Baseline: And we're going to do that on your hair today.</p> <p>Our model: And we're going to do that on her hair today.</p>
	<p>Source: Ela pegará neve e a empurrará para o lado da estrada ou ela pegará a sujeira de um ponto alto e a moverá para o lado.</p> <p>Reference: It will catch snow and push it over to the side of the road or it will catch dirt out of a high spot and move it over to the side.</p> <p>Baseline: She will take snow and push it to the side of the road or she will take the dirt from a high point and move it to the side.</p> <p>Our model: It will take snow and push it to the side of the road or it will take the dirt from a high spot and move it to the side.</p>

Table 2: Translation examples of random masking VTLM (baseline) and more informed visual masking VTLM (our model).

performance of our model was superior to the baseline by approximately 0.7 METEOR. However, in terms of BLEU, the performance was inferior to the baseline by approximately 0.16 BLEU, presenting a behavior different from that observed in the last two experiments.

Therefore, the results indicate that more informed visual masking benefits the final performance of the model to a certain extent. By increasing the proportion of T_{People} tokens being masked, there is an improvement in the performance of the model compared to the baseline. Nevertheless, when this proportion becomes greater than 50%, this improvement tends to decrease. This behavior may be explained by the decrease in tokens related to other categories being masked since the visual masking ratio did not change, i.e., it remained at 15%. Thus, excessively increasing the proportion of T_{People} tokens being masked can jeopardize the learning of elements from other categories.

Qualitative Analysis. To better understand the effect of our proposed pre-training masking approach, we compare some examples of texts translated by random masking VTLM (baseline) and more informed visual masking VTLM (our model). The

examples are presented in Table 2. In the first example, the baseline mistranslates the subject pronouns “he” and “she”, translating both to “it”, while our model translates them correctly, achieving better performance. In the second example, the baseline mistranslates the object pronoun “him”, translating it to “it”, while our model translates it correctly. The third example illustrates the correct translation of the possessive adjective “her” by our model, while the baseline mistranslates it to “your”. Finally, the baseline references an object using the subject pronoun “she” instead of “it”. In contrast, our model does not make the same mistake and uses the pronoun correctly.

4.2 More informed textual masking

We run the same experiment using three different ratios of T_{Pronouns} – 33.34%, 50.0%, and 66.67% – and the results are shown in Table 3. The results show that this masking strategy also affects the final performance of the model. For $T_{\text{Pronouns}} = 33.34\%$, our model scored 52.64 BLEU and 79.45 METEOR on the test set and 52.96 BLEU and 79.53 METEOR on the valid set, outperforming the baseline by approximately 0.7 BLEU and 1.3

Model	T_{Pronouns}	Test		Valid	
		BLEU	METEOR	BLEU	METEOR
VTLM: random masking		51.80	78.04	52.44	78.25
	33.34%	52.64	79.45	52.96	79.53
	50.00%	52.39	79.35	52.94	79.51
VTLM: more informed textual masking	66.67%	52.21	79.27	52.82	79.42

Table 3: BLEU and METEOR scores for random masking VTLM (baseline) and more informed textual masking VTLM (our model) for the MMT task.





	Source: Se você andar seu cachorro do seu lado esquerdo, você quer que ele se sente do lado, porque o que ele faz é apertar, então, se você estiver por aqui, o cachorro deveria tê-lo aqui.
	Reference: If you walk your dog on your left side you want it to sit on the side because what it does is tighten up so if you're over here the dog should have it over here.
	Baseline: If you walk your dog on your left side you want him to sit on the side because what he does is squeeze, then if you're standing over here the dog should have him here.
	Our model: If you walk your dog on your left side you want it to sit on the side because what it does is tighten, then if you're over here the dog should have it here.
	Source: Ela entra em cena depois que a cena começa entre o policial e Stanley.
	Reference: She walks into the scene after the scene begins between the police officer and Stanley.
	Baseline: It goes into scene after the scene starts between the police officer and Stanley.
	Our model: She goes into scene after the scene starts between the police officer and Stanley.
	Source: E eu só trabalhei uma noite com ela.
	Reference: And I only worked one night with her .
	Baseline: And I just worked a night with it .
	Our model: And I just worked a night with her .
	Source: Mas, eu vou tentar de qualquer maneira e você pode ter uma ideia do que você pode querer fazer.
	Reference: But, I'm going to try it anyway and you can get an idea of what you might want to do.
	Baseline: But, I'm going to try anyway and you might have an idea of what you might want to do.
	Our model: But, I'm going to try it anyway and you might get an idea of what you might want to do.

Table 4: Translation examples of random masking VTLM (baseline) and more informed textual masking VTLM (our model).

METEOR. As for $T_{\text{Pronouns}} = 50.0\%$, our model also surpassed the baseline, but its performance was worse than in the previous experiment. Finally, for $T_{\text{Pronouns}} = 66.67\%$, our model performed better than the baseline in terms of BLEU and METEOR, but its performance was inferior than in the last two experiments, when the chosen proportions were 33.34% and 50.0%.

Therefore, the results indicate that masking more T_{Pronouns} tokens leads to an improvement in the final performance of the model. However, even though our model surpassed the baseline in all experiments, this performance improvement is limited, as the best performance was observed when T_{Pronouns} proportion was 33.34%, followed by 50.0% and 66.67%, respectively.

Qualitative Analysis. Some examples of texts translated by each model are presented in Table 4.

In the first example, the random masking VTLM uses the pronouns “he” and “him” to refer to the word “dog” instead of using the pronoun “it”, which should have been used in this case. On the other hand, our model does not make the same mistake and uses the correct pronoun in all cases, achieving better translation performance. In the second example, the random masking VTLM mistranslates the subject pronoun “she” and translates it to “it”, which is a serious translation error since the pronoun “it” cannot be used to refer to a person. In contrast, our model uses the correct pronoun and achieves better performance. The next example illustrates the incorrect translation of the object pronoun “her” by the baseline, which again uses the pronoun “it” to refer to a person. However, this error is not made by our model, which makes the correct use of the pronoun in the translation.

The three previous examples illustrate situations similar to those observed with the application of more informed visual masking. However, the last example shows a further improvement in translation. This improvement is related to the use of the pronoun “it” as the direct object of a verb. While the baseline omits this pronoun in the translation, our model correctly uses it after the verb “try”.

4.3 More informed visual and textual masking

Table 5 shows BLEU and METEOR scores across valid and test sets of How2. The obtained results show that the more informed visual and textual masking strategy also affects the performance of the MMT model. Our model achieved 52.34 BLEU and 78.77 METEOR on the test set and 53.28 BLEU and 79.44 METEOR on the valid set, outperforming the baseline by approximately 0.7 BLEU and 0.9 METEOR.

Although the performance improvement was not very high in terms of BLEU and METEOR, the results indicate that applying more informed visual and textual masking benefits the final performance of the model.

Qualitative Analysis. To further understand the effectiveness of our approach, we compared some examples of texts translated by random masking VTLM (baseline) and more informed visual and textual masking VTLM (our model). The examples are presented in Table 6. In the first example, the random masking VTLM references the word “website” using the subject pronoun “he” instead of the pronoun “it”. In contrast, our model does not make the same mistake and uses this pronoun correctly. In the second example, the object pronoun “him” is used incorrectly by the baseline. In this case, the pronoun “it” should have been used and our model makes the correct use of this pronoun. The third case illustrates the correct translation of the possessive adjective “your” by our model, while the baseline mistranslates it to “their”. In the fourth example, our model correctly uses the pronoun “it” as the direct object of the verb “take”, while the baseline omits this pronoun in the translation.

Finally, the last situation illustrates a new improvement not seen when applying more informed visual masking or more informed textual masking separately. Although visual information improves the overall performance of the standard multimodal model, we observed that it can lead to the incorrect use of certain pronouns. For instance, when

the video frame associated with the text has an element categorized as “man”, the pronouns used in the translation tend to be “he” or “him”. Likewise, when there is an element categorized as “woman” in the video frame, the pronouns tend to be “she” or “her”. On the other hand, our more informed masking approach tends to better deal with this bad tendency of multimodal models. In the last example, the two elements categorized as “man” in the image possibly influenced the incorrect choice of the pronoun “him” after the verb “bring” by the baseline model. However, our model did not make the same mistake and used the pronoun “it” correctly.

5 Related Work

Pre-trained language models have become essential in the natural language processing field. One pre-trained model that has attracted considerable attention in this field is BERT (Devlin et al., 2019). BERT introduces masked language modeling (MLM) to efficiently learn bidirectional representations by masking a set of input tokens at random and predicting them afterward. In this approach, 15% of input tokens are randomly selected for masking, from which 80% are replaced with the [MASK] token, 10% are replaced with a random token, and 10% are left intact.

Following BERT, several approaches have been proposed to optimize pre-trained language models. Devlin et al. (2019) later propose whole word masking (wwm) in an attempt to address the drawbacks of random token masking in the MLM task. In this approach, input tokens are segmented into units corresponding to whole words, and instead of selecting tokens to mask at random, they mask all of the tokens corresponding to a whole word at once. Zhang et al. (2019) introduce ERNIE to optimize the masking process of BERT by applying entity/phrase masking. Instead of randomly selecting input words, phrase-level masking masks consecutive words and entity-level masking masks the named entities. Clark et al. (2020) present ELECTRA, which uses a generator-discriminator framework. While the generator learns to predict the original words of the masked tokens, the discriminator uses Replaced Token Detection to discriminate whether the input token is replaced by the generator. Levine et al. (2021) propose a principled masking strategy based on the concept of Pointwise Mutual Information (PMI). PMI-masking jointly

Model	Test		Valid	
	BLEU	METEOR	BLEU	METEOR
VTLM: random masking	51.80	78.04	52.44	78.25
VTLM: more informed visual and textual masking	52.34	78.77	53.28	79.44

Table 5: BLEU and METEOR scores for random masking VTLM (baseline) and more informed visual and textual masking VTLM (our model) for the MMT task.






	Source: Isso é o que dá ao meu site as opções de cores que ele tem. Reference: That is what gives my site the color options that it has. Baseline: That's what gives my website to the color options that he has. Our model: That's what gives my website to the color options that it has.
	Source: E eu vou empurrá-lo de volta. Reference: And I'm going to push it back down. Baseline: And I'm going to push him back. Our model: And I'm going to push it back down.
	Source: Eles mantêm seus dedos juntos e são bons para muitas atividades. Reference: They keep your fingers kind of together and are good for a lot of activities. Baseline: They keep their fingers together and they're good for many activities. Our model: They keep your fingers together and they're good for a lot of activities.
	Source: Agora pegue, coloque a marca do oleiro lá. Reference: Now take it , put the potter's mark in there. Baseline: Now take, put your potter's mark on there. Our model: Now take it , put the potter's mark in there.
	Source: Ele quer trazê-lo de volta naturalmente. Reference: He wants to bring it back naturally. Baseline: He wants to bring him back naturally. Our model: He wants to bring it back naturally.

Table 6: Translation examples of random masking VTLM (baseline) and more informed visual and textual masking VTLM (our model).

masks a token n -gram if it exhibits high collocation over the corpus.

Combining cross-lingual and visual pre-training, Caglayan et al. (2021) propose Visual Translation Language Modelling (VTLM), which extends the TLM framework (Conneau and Lample, 2019) with regional features and performs masked language modeling and masked region classification on a three-way parallel language and vision dataset. The standard masking ratio is maintained (i.e. 15%) and it is applied separately to visual and language flows. VTLM achieved a 44.0 BLEU and 61.3 METEOR on the English-German 2016 test set of Multi30k (Elliott et al., 2016) for the MMT task. Following this approach, Sato et al. (2022) propose VTLM for video subtitles, which extends VTLM to a new language pair and to more challenging circumstances concerning the image-text relationship by using video frames with subtitles instead of images with their corresponding description. They use the same

random masking approach for both visual and textual masking and achieved a 51.8 BLEU and 78.0 METEOR on the Portuguese-English test set of How2 (Sanabria et al., 2018) for the MMT task. In this paper, we propose three novel masking strategies for cross-lingual visual pre-training and we apply them to VTLM for video subtitles to test their efficacy for downstream MMT performance.

6 Conclusions

In this work, we show that predicting particular masked elements can benefit cross-lingual visual pre-training as the pre-trained model can acquire a better understanding of specific language structures, which improves downstream tasks such as multimodal machine translation. We present three selective masking strategies that focus on masking specific linguistic and visual tokens that can contribute to understanding some language patterns.

We achieve state-of-the-art accuracy on the How2 dataset and show that our masking approaches yield significant improvements over the original random masking strategy for downstream MMT performance. Even though we only conduct experiments on the MMT task using VTLM as the base model, our method can easily generalize to other models and other NLP tasks. We hope that our work here will further accelerate future research on Brazilian Portuguese and other low-resource languages. For future work, we will investigate the impact of visual and textual masking probability and further explore more effective masking approaches for downstream MMT performance.

Limitations

Although our research led to improvements in the translation of subject pronouns, object pronouns, and possessive adjectives and pronouns, these improvements did not cover non-binary-associated pronouns, such as *they/them/theirs*, *xe/xem/xyr* and *ze/hir/hirs*. The large underrepresentation of non-binary genders in textual and visual data contributes to propagating the misrepresentation of non-binary people by language models. In this paper, we were unable to work against this issue, thus we hope to contribute to a fairer representation of these disadvantaged groups in the future.

Ethics Statement

We acknowledge that all co-authors of this paper are aware of the *ACM Code of Ethics* and honor the code of conduct. We collected our data from a public dataset that permits academic use. As our experiments are limited to the binary linguistic forms represented in the used data, we cannot guarantee that our models will always generate unbiased content.

References

Alexei Baevski, Sergey Edunov, Yinhan Liu, Luke Zettlemoyer, and Michael Auli. 2019. [Cloze-driven pretraining of self-attention networks](#). 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 5360–5369, Hong Kong, China. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with im-](#)

[proved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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, Belgium, Brussels. Association for Computational Linguistics.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. [Cross-lingual Visual Pre-training for Multimodal Machine Translation](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Short Papers*, online. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. [UNITER: Universal image-text representation learning](#).

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [ELECTRA: Pre-training text encoders as discriminators rather than generators](#). In *ICLR*.

Alexis Conneau and Guillaume Lample. 2019. [Cross-lingual language model pretraining](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

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*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. [Multi30K: Multilingual English-German image descriptions](#). In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. [SpanBERT: Improving pre-training by representing and predicting spans](#). *Transactions of the Association for Computational Linguistics*, 8:64–77.
- Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.
- Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, and et al. 2020. [The open images dataset v4](#). *International Journal of Computer Vision*, 128(7):1956–1981.
- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. [ALBERT: A lite BERT for self-supervised learning of language representations](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2021. [{PMI}-masking: Principled masking of correlated spans](#). In *International Conference on Learning Representations*.
- 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](#). *CoRR*, abs/1907.11692.
- 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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. [Faster R-CNN: Towards real-time object detection with region proposal networks](#). In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.
- Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loic Barrault, Lucia Specia, and Florian Metze. 2018. [How2: A large-scale dataset for multimodal language understanding](#). In *Visually Grounded Interaction and Language (ViGIL), Montreal; Canada, December 2018. Neural Information Processing Society (NeurIPS)*, arXiv. arxiv.org. 32nd Annual Conference on Neural Information Processing Systems, NeurIPS ; Conference date: 02-12-2018 Through 08-12-2018.
- Júlia Sato, Helena Caseli, and Lucia Specia. 2022. [Multilingual and multimodal learning for Brazilian Portuguese](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 919–927, Marseille, France. European Language Resources Association.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [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.
- Lucia Specia, Stella Frank, Khalil Sima’an, and Desmond Elliott. 2016. [A shared task on multimodal machine translation and crosslingual image description](#). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.
- Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. [Dropout: A simple way to prevent neural networks from overfitting](#). *Journal of Machine Learning Research*, 15(56):1929–1958.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. [VL-BERT: Pre-training of generic visual-linguistic representations](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. [Xlnet: Generalized autoregressive pretraining for language understanding](#). In *Neural Information Processing Systems*.
- Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. [ERNIE: Enhanced language representation with informative entities](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.
- Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. [A robustly optimized BERT pre-training approach with post-training](#). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.