Improving Vision-Language Cross-Lingual Transfer with Scheduled Unfreezing

Max Reinhardt[1] **Gregor Geigle**[1,2] **Radu Timofte**[2] **Goran Glavaš**[1]
[1]WüNLP, [2]Computer Vision Lab, CAIDAS, University of Würzburg,
gregor.geigle@uni-wuerzburg.de

Abstract

Large-scale pretraining of vision-language (VL) models brought dramatic improvements across numerous tasks, from visual question-answering to cross-modal retrieval but these gains are mostly limited to English. Massively multilingual VL encoder models (mVLMs) hold promise for other languages: after fine-tuning on only English task data, they can perform the task in other languages in what is termed zero-shot cross-lingual transfer (ZS-XLT). Still, ZS-XLT sees a large performance gap to English, especially for low-resource languages. In this work, we reduce this gap with a fine-tuning strategy known as *Scheduled Unfreezing* (SUF): instead of updating all parameters from the start, we begin with the top layer(s) of the vision-language encoder and gradually unfreeze (i.e., update) its layers top to bottom. SUF forces reliance on encoder's representations from higher layers: the fact that in multilingual models these representations encode higher-level semantics rather than low-level language-specific idiosyncrasies, we hypothesize, should render SUF beneficial for ZS-XLT. Experiments with two mVLMs (UC2 & CCLM) on three downstream tasks (xGQA, XVNLI, xFlickrCo) show that SUF brings consistent gains in ZS-XLT, especially for visual Q&A (xGQA) by up to 10 points.

1 Introduction

Recent vision-language (VL) models (Zhou et al., 2021; Zeng et al., 2022; Li et al., 2023a; Liu et al., 2023c; Geigle et al., 2023, inter alia), trained on massive amounts of image-text data, led to dramatic improvements on virtually all VL tasks (e.g., image-text retrieval or visual Q&A). This progress, however, benefits primarily English. Large Vision-Language models (LVLMs) (Li et al., 2023a; Liu et al., 2023c,b; Dai et al., 2023; Bai et al., 2023)—which align an image encoder to a Large Language Model (LLM)—excel in generalizing *zero-shot* to new tasks (without task-specific fine-tuning). Most LVLMs use English LLMs and are not highly multilingual; they fail to follow instructions in other languages or produce English output (Geigle et al., 2023; Kew et al., 2023; Holtermann et al., 2024; Shaham et al., 2024). Multilingual LVLMs are much less available[1] and generally underperform their English counterparts (Geigle et al., 2023).

The alternative is task-specific fine-tuning of smaller, but massively multilingually pretrained VL *encoder* models (mVLMs) (Ni et al., 2021; Zhou et al., 2021; Zeng et al., 2022). Here, however, task-specific training data exists predominantly in English which forces us to rely on *zero-shot cross-lingual transfer* (ZS-XLT) (Conneau et al., 2020b; Lauscher et al., 2020): due to the massively multilingual pretraining, the encoders fine-tuned on English task data can be used for inference in other languages. Still, ZS-XLT results in substantial performance drops in other languages compared to English, especially for less represented target languages in m(V)LM's pretraining. While few-shot training for specific target languages can reduce this performance gap (Lauscher et al., 2020; Schmidt et al., 2022), annotating sufficient data (for training and model validation) is expensive and does not scale to hundreds of languages.

In this work, we improve ZS-XLT with mVLMs using a training method known as *scheduled unfreezing* (SUF) (Howard and Ruder, 2018a; Liu et al., 2024). SUF, which we apply in task-specific fine-tuning of an mVLM on English data, gradually increases the set of encoder's (i.e., Transformer's) parameters that are being fine-tuned (i.e., updated), starting from the last layer(s) and gradually adding lower layers of the Transformer stack as the training progresses. Multilingual language-only encoders have been shown to encode language-agnostic high-level semantic knowledge in higher layers and language-specific idiosyn-

[1]Powerful multilingual LVLMs such as Google's PaLI models (Chen et al., 2023) are, unfortunately, not public.

Proceedings of the 3rd Workshop on Advances in Language and Vision Research, pages 155–166
August 16, 2024 ©2024 Association for Computational Linguistics

crasies in lower layers (Libovický et al., 2020; Hu et al., 2020). If the same holds for mVLMs, then SUF—by enforcing stronger reliance on representations from higher layers of an mVLM—should facilitate ZS-XLT for VL tasks. Put differently, with SUF fine-tuning on English-only data, idiosyncratic English-specific knowledge from lower layers of the encoder is less available, forcing the model to rely on more language-agnostic knowledge from higher layers of the encoder.

We evaluate the effects of SUF fine-tuning on ZS-XLT for two multilingual vision-language encoders: UC2 (Zhou et al., 2021) and CCLM (Zeng et al., 2022); and on three distinct downstream tasks: visual QA (xGQA (Pfeiffer et al., 2022)), image-text retrieval (xFlickrCo (Bugliarello et al., 2022)), and visual entailment (XVNLI (Bugliarello et al., 2022)). We find that SUF consistently improves performance compared to standard fine-tuning: by up to 3 points in retrieval and entailment and by a massive 10 points for visual QA.

Our further fine-grained analysis of model behavior on xGQA reveals that: (1) in standard fine-tuning the performance for most target languages stagnates or degrades over the course of (English) training, while the English performance steadily improves. (2) in SUF fine-tuning, in contrast, trajectory of target language performance longer mirrors that of English performance, suggesting that the model relies on more language-agnostic representations; this results in massive improvements especially for some languages distant from English, such as Korean and Bengali. Using parallel data, we show that SUF fine-tuning indeed leads to cross-lingually more aligned representations of the sequence start token ([CLS]), which is input to the classifier. Finally, we compare SUF against two other strategies that similarly reduce reliance on lower layers of the encoder: (1) layer-wise learning rate decay and (2) fixed training of only the top layers. While both these also yield some performance gains, they underperform SUF. SUF-based fine-tuning not only improves ZS-XLT of mVLMs but is also computationally more efficient than standard fine-tuning: we thus hope that our work motivates broader investigation of SUF strategies in the context of multilingual VL models.

2 Related Work

Cross-lingual Transfer with Vision-Language Models. Bugliarello et al. (2022) created the IGLUE benchmark, which has become the de facto benchmark for evaluating cross-lingual transfer abilities of mVLMs. IGLUE comprises four VL tasks: visual QA (xGQA (Pfeiffer et al., 2022)), visual entailment (XVNLI (Xie et al., 2019)), multi-image reasoning (MaRVL) (Suhr et al., 2019; Liu et al., 2021a), and image-text retrieval (Lin et al., 2014; Plummer et al., 2015). Being designed specifically for ZS-XLT, each dataset in IGLUE comes with a training portion in English and test portions in different target languages.

Bugliarello et al. (2022) compare several multilingual VL encoder models on IGLUE, namely: M3P (Ni et al., 2021), x/mUNITER (Liu et al., 2021a), and UC2 (Zhou et al., 2021)), primarily in ZS-XLT, but also in few-shot cross-lingual transfer (FS-XLT) in which few training instances in target languages are assumed to exist. Crucially, in both setups they demonstrate significant gaps between models' English performance and their performance for other languages. Subsequent models such as CCLM (Zeng et al., 2022), Li et al. (2023b), and Ernie-UniX2 (Shan et al., 2022) improved target-language performance, but since their English performance improved as well, this resulted overall in similar ZS-XLT performance gaps.

For visual question answering in particular, there has been work dedicated to reducing the cross-lingual performance gap. Nooralahzadeh and Sennrich (2023) assessed that a high ambiguity in the label space makes learning more difficult, attempting to remedy for this with several strategies, including addition of a similarity-based loss to standard classification cross-entropy loss, code-switching at the instance level and a sparse fine-tuning approach. Liu et al. (2023a) reduce the ZS-XLT performance gap by replacing the standard single-layer classifier with a deeper two-layer architecture. Observing stark performance differences across different question types, they also introduced a special question-type token.

Finally, Geigle et al. (2023) find that fine-tuning a multilingual LVLM that relies on mT0 (Xue et al., 2021; Muennighoff et al., 2022) as the LLM backbone nearly closes the ZS-XLT gap. Training and fine-tuning billion-parameter LVLMs is, however, much more computationally expensive; crucially, the same is true for inference, which hinders model application for most users. Moreover, Geigle et al. (2023) show that the cross-lingual performance gap is highly dependent on the backbone LLM, ob-

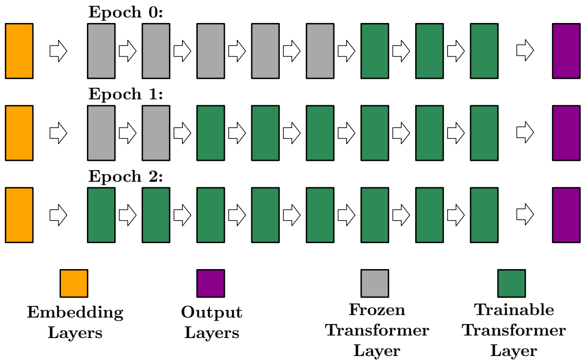


Figure 1: Illustration of Scheduled Unfreezing; each rectangle shows one Transformer layer, green rectangles denote unfrozen layers whereas gray ones indicate frozen layers. The embedding layer (orange) is kept unfrozen along with the task-specific classification head (purple). In every epoch, we unfreeze a fixed number of layers from top to bottom.

serving larger ZS-XLT gaps with BLOOMZ (Scao et al., 2022; Muennighoff et al., 2022).

In this work, we focus on encoder mVLMs, due to their smaller computational footprint and thus broader applicability. To the best of our knowledge, our SUF is the first strategy shown to substantially reduce the ZS-XLT gap for VL encoders.

Unfreezing training strategies. Various strategies for (un)freezing model parts have been proposed in transfer learning scenarios. Howard and Ruder (2018b) introduce Gradual Unfreezing for fine-tuning a pretrained recurrent LM, to avoid catastrophic forgetting across different text classification tasks; in each epoch, starting from the top layer, they unfreeze one layer of the pretrained LM. However, Raffel et al. (2020) find that this underperforms full model fine-tuning for Transformer-based LMs. In the context of XLT with multilingual LMs, in concurrent work Liu et al. (2024) propose a scoring function that dynamically decides when and which layers to unfreeze. In this work, in contrast, we investigate a simpler fixed unfreezing schedule and focus on bimodal vision-language models rather than unimodal language-only models.

3 Scheduled Unfreezing

The exact setup on which we focus in this work is zero-shot cross-lingual transfer (ZS-XLT) for downstream vision-language tasks (e.g., visual QA) with massively multilingual vision-language encoder models (mVLMs) as vehicles of the transfer. In this setup, we fine-tune the mVLM on task-specific data in English only and evaluate its performance on task-specific data in other languages.

Based on the observation (from multilingual language-only encoders) that multilingual encoders encode more language-agnostic higher-order semantics in their upper Transformer layers and language-specific information in their lower layers (Libovický et al., 2020; Hu et al., 2020), we propose fine-tuning based on top-to-bottom **scheduled unfreezing** (SUF) as a method to facilitate cross-lingual transfer with mVLMs. The motivation for SUF in this context is as follows: by (initially) freezing lower Transformer layers, the classification head is forced to solve the task by tuning language-agnostic knowledge from higher Transformer layers of the mVLM first. Contrary, in full fine-tuning, the classifier can additionally leverage language-specific knowledge from lower layers—when fine-tuned on English tasks data only. This means that the classifier is more likely to overfit to English-specific features, harming the effectiveness of cross-lingual transfer to other languages.

To test this hypothesis, we use a fixed-schedule unfreezing in this work, illustrated in Figure 1. The general idea is not to train the full model from the start, but freeze (i.e., not update) all but the top k layers at the beginning and then gradually unfreeze k layers top-to-bottom in every epoch.

Architecture-specific Implementation. Compared to unimodal language-only encoders (Devlin et al., 2019; Conneau et al., 2020b), mVLMs additionally contain components for encoding the visual modality (i.e., images). Moreover, mVLMs come with different architectures, differing primarily w.r.t. where cross-modal information aggregation occurs. As such, we introduce architecture-specific unfreezing schedules for the two mVLMs with which we experiment in this work: UC2 (Zhou et al., 2021) and CCLM (Zeng et al., 2022).

UC2 is an encoder Transformer model, architecturally identical to the language-only XLM-R encoder (Conneau et al., 2020a). UC2 encodes an image offline, relying on an object detection model (Ren et al., 2015)[2]; the features for image regions given by this model are linearly projected and then concatenated with the text embeddings as input to the model. The image region vectors are treated by the Transformer like any other text token. As a result, we can use general SUF without any adjust-

[2]All images are processed prior to training and the detection model is not used during training of UC2.

ments: UC2, using a base-size XLM-R architecture, has 12 Transformer layers. In the first epoch, the task-specific classification head, the embedding layer[3], and the top $k = 3$ Transformer layers remain unfrozen. After every training epoch, we unfreeze 3 additional layers, top to bottom.

CCLM, also a Transformer-based encoder, comprises n layers for processing only the text input, followed by m more cross-modal layers, which additionally have a cross-attention component. Through this cross-attention, the model attends to the image features extracted by a separate Vision Transformer (ViT) (Dosovitskiy et al., 2020). For $\text{CCLM}_{\text{base}}$, which we use in our experiments, there are n=12 layers for pure text encoding (initialized from XLM-R), followed by m=6 cross-modal layers (initialized from X2-VLM (Zeng et al., 2023)). We keep the ViT fully unfrozen during training. The motivation for this is twofold: (i) the resolution of images in fine-tuning is larger (384x384) than in its pretraining (224x224), requiring ViT to adapt; and (ii) we employ SUF to reduce the impact of language-specific (i.e., English) overfitting in fine-tuning and image encoding with ViT is inherently language-agnostic. We thus keep the ViT, task-specific classification head, and embedding layer unfrozen throughout training. In the first epoch, we additionally start with the top $k = 3$ Transformer layers (out of $m + n$=18) unfrozen and then unfreeze 3 more layers after each epoch.

4 Evaluation

We provide details of our experimental setup and then consider results over three downstream tasks with the two architectures (UC2 & CCLM).

4.1 Experimental Setup

Datasets. We evaluate SUF on the multilingual IGLUE benchmark (Bugliarello et al., 2022) for ZS-XLT. IGLUE spans 4 different tasks: visual QA (**xGQA** (Pfeiffer et al., 2022; Hudson and Manning, 2019)), image-text retrieval (**xFlickrCo** (Bugliarello et al., 2022)), visual entailment (**XVNLI**) (Xie et al., 2019; Bugliarello et al., 2022), and multi-image reasoning (**MaRVL** (Liu et al., 2021b)). We exclude MaRVL, because it requires changes to the model architecture in order to support multi-image input.

[3]Initial experiments showed that keeping the embedding layer unfrozen was critical for good performance.

xGQA contains diverse questions over multiple question types – Verify (yes/no), Query (open), Choose (one of two options), Logical (true or false), Compare (across multiple objects) – with nearly 2000 unique labels. This dataset is obtained by extending the monolingual GQA (Hudson and Manning, 2019) with human translations in 7 languages. The English training portion contains 943K examples. We report classification accuracy.

For image-text retrieval, the task is to retrieve the best caption for an image (Text Retrieval, TR) or the corresponding image given a caption (Image Retrieval, IR). We use xFlickrCo which couples 1K images from Flickr30K (Plummer et al., 2015) test portion with 1K images from the COCO (Lin et al., 2014) test portion with human-written captions in 7 languages (plus the original English Flickr30k and MSCOCO captions). For training, we use the Flickr30k training split with 145K examples. As metric, we report recall@1 (R@1)—the proportion of images (in TR) or captions (in IR) for which the matching caption (in TR) or image (in IR) is positioned at the very top of the ranking.

For visual entailment on XVNLI, a model must predict if a statement (i.e., a hypothesis), is entailed, contradicts, or is neutral to an image (as the "premise"). The training portion of the dataset consists of 541K English examples and the test portion covers 4 other languages (Arabic, Spanish, French, and Russian). We report results in terms of classification accuracy.

Training Setup. We mirror the training procedures from IGLUE and (Zeng et al., 2022) for task-specific fine-tuning of of UC2 and CCLM. For xGQA with UC2, we add a 2-layer classification head (with $\sim$ 2000 classes, i.e., valid answers from the training data). CCLM casts VQA as a generation task, adding a full-blown 6-layer decoder Transformer (the input to which is the representation of the `[CLS]` token, output of the last layer of the CCLM's cross-encoder). The decoder Transformer is trained on the task and as such not frozen.

Hyperparameters: We train for the same number of epochs for each task as in IGLUE: 5/10/10 epochs, for xGQA, XVNLI, and xFlickrCo, respectively. Regarding other hyperparameter values, we follow IGLUE for training UC2, using the learning rate of $4 \cdot 10^{-5}$ for xGQA and $2 \cdot 10^{-5}$ for XVNLI and xFlickrCo. We train in batches of size 256 for xGQA, 64 for xFlickrCo, and 128 for XVNLI. For CCLM (the original work did not report fine-tuning hyperparameter values), we use a learning rate of $2 \cdot 10^{-5}$ for the image encoder (i.e., ViT) and $3 \cdot 10^{-5}$ for the rest of the model. We use an effective batch size of 256/128/144 for xGQA, xFlickrCo, and XVNLI, respectively, resorting to gradient accumulation, due to limited GPU VRAM[4]. For both models and in all fine-tuning procedures, we use AdamW (Loshchilov and Hutter, 2019) optimizer, with linear warm-up for 10% of steps and weight decay of 0.01. We use exactly the same hyperparameters for standard and SUF fine-tuning.

Evaluation Setup. We compare SUF fine-tuning against standard full fine-tuning for ZS-XLT. In other words, we fine-tune the model on the task-specific English training split and then evaluate its performance on the same task on the test splits in English and other languages. We evaluate all models, with and without SUF, after the last training epoch. For xFlickrCo, with CCLM, we first pre-filter 128 best image (in IR) or captions (in TR) matches based on the cosine similarity of their image and text representations (computed independently from the other modality using the image encoder and the text-only layers), and then re-rank the candidates by jointly scoring all candidates. With UC2, we directly compute the pairwise similarity of all possible image-text pairs. For xGQA with CCLM, we perform constrained generation to the set of task-specific class labels.

4.2 Results

The overview of the ZS-XLT results (together with English performance), aggregated over all target languages for each task, is given in Table 1. Scheduled unfreezing (SUF) yields consistent ZS-XLT performance gains over standard fine-tuning for all three tasks and both UC2 and CCLM. At the same time, the English performance in SUF is comparable to that of standard fine-tuning. This means that not only does (1) SUF fine-tuning truly reduce the cross-lingual performance gap for mVLMs, but (2) freezing of lower layers does not seem to hurt the source language performance. While SUF fine-tuning of CCLM brings moderate 2-3 point improvements on XVNLI and xFlickrCo, on xGQA we observe a massive 10-point average gain over the 7 target languages. We next investigate the xQGA performance in more detail.

[4]For xFlickrCo, where we use in-batch negatives, this yields lower scores than reported in Zeng et al. (2022).

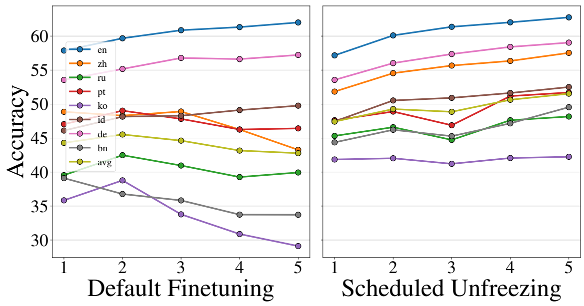


Figure 2: Results on xGQA for $\text{CCLM}_{\text{base}}$ after each epoch for each language. We compare the standard finetuning (left) with scheduled unfreezing (SUF) fine-tuning (right).

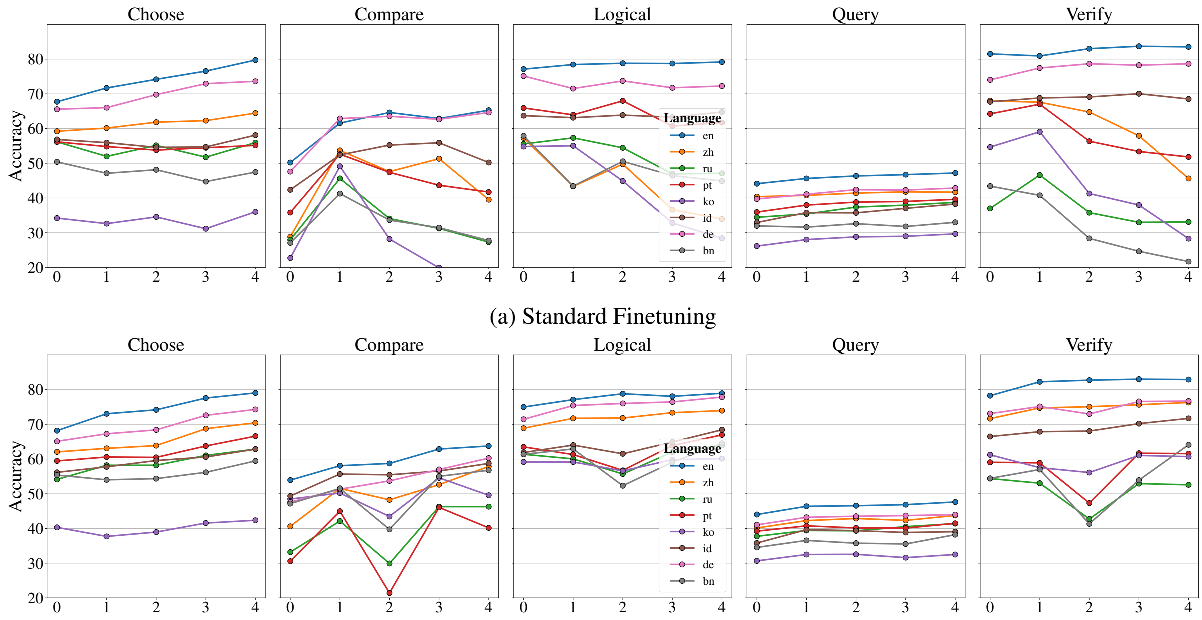
In-Depth Analysis for xGQA. Motivated by the large performance gains that SUF fine-tuning brings in ZS-XLT for xGQA, we next inspect model behavior on this task in more detail, across two performance dimensions: (i) individual target languages and (ii) different question types, aiming to unravel factors that specifically contribute to good ZS-XLT performance.

Per-Language Performance. We first analyze how training on English data affects the transfer to other languages for different training duration. In Figure 2, we show the per-epoch accuracy of $CCLM$ for all target languages (and EN as the source language. With standard fine-tuning, English performance improves throughout the training; the performance for most other languages, however, either stagnate or decreases. The only exception to this pattern is German (DE), which is not only a high-resource language but also linguistically closest to English. For languages most distant from English, Korean and Bengali, we observe largest performance drops with prolonged English training. Scheduled unfreezing, on the other hand, prevents this performance decay and most languages benefit from longer English training under SUF fine-tuning. Additionally, we see that most languages also start at a higher accuracy with scheduled unfreezing. This suggests that the freezing of lower layers at the start forces the model to rely on more language-agnostic features that transfer better.

Per-Question Type Performance. GQA is constructed around 5 question types: *Verify* (yes/no), *Query* (open), *Choose* (one out of two options), *Logical* (true or false), and *Compare* (across multiple objects). Figure 3 summarizes the ZS-XLT performance for different question types across the training epochs. We see that SUF fine-tuning pre-

Setup	xGQA		XVNLI		xFlickrCo			
					TR		IR	
	EN	ZS-XLT	EN	ZS-XLT	EN	ZS-XLT	EN	ZS-XLT
UC2	57.1	31.9	77.1	**61.7**	**36.8**	18.0	**43.0**	20.0
UC2+SUF	57.1	**41.3**	77.2	61.2	36.4	**20.0**	41.8	**22.3**
CCLM	62.0	42.8	**81.2**	68.6	77.7	63.4	78.0	64.2
CCLM+SUF	**62.8**	**51.5**	80.6	**70.6**	**78.5**	**66.7**	**78.6**	**67.1**

Table 1: Evaluation of SUF on UC2 and $\text{CCLM}_{\text{base}}$ across multiple V&L tasks. We report results for English (en) and averaged (avg) across all non-English languages. We **bold** the best results. We report accuracy for xGQA and XVNLI, and recall@1 for xFlickrCo for both Text Retrieval (TR) and Image Retrieval (IR).



(a) Standard Finetuning

(b) Scheduled Unfreezing

Figure 3: Accuracy every epoch for each question type in xGQA for SUF and standard fine-tuning with $\text{CCLM}_{\text{base}}$.

vents language-specific overfitting to English in particular for *Compare*, *Logical*, and *Verify* questions. It is worth noting that all three question types effectively have only 'yes' and 'no' as answer labels. This means that SUF is not improving ZS-XLT by reducing label space ambiguity (like, e.g., Nooralahzadeh and Sennrich (2023)), but rather by preventing early overfitting to English-specific idiosyncrasies in the questions.

Expectedly, all models generally exhibit the lowest performance on the open-ended *Query* questions, which account for the largest portion of the xQGA data. For both *Query* and *Choose* questions, English training with both standard and SUF fine-tuning generally increases the performance for target languages throughout the training; for SUF fine-tuning, however, the starting accuracy scores are higher than for standard fine-tuning, resulting in overall better scores at the end of training.

5 Further Analysis

We further analyze SUF fine-tuning through the lens of cross-language similarity of [CLS] tokens for parallel data. We then compare SUF with conceptually similar alternatives: (i) layer-wise learning rate decay and (ii) updating only the top layers Transformer layers throughout the whole training. Finally, we report the results of few-shot cross-lingual transfer (FS-XLT).

5.1 Cross-Lingual Semantic Alignment

Our previous findings suggest that SUF can retain the cross-lingual transfer abilities of the mVLM better than standard finetuning. We thus further test cross-lingual semantic alignment for both fine-tuning regimes (with UC2), using parallel data.

SF	bn	de	en	id	ko	pt	ru	zh
bn	100	45	33	48	58	47	55	48
de	45	100	61	58	48	55	60	57
en	33	61	100	53	39	49	52	56
id	48	58	53	100	50	57	60	60
ko	58	48	39	50	100	54	57	56
pt	47	55	49	57	54	100	58	56
ru	55	60	52	60	57	58	100	60
zh	48	57	56	60	56	56	60	100

(a) xGQA: Standard Finetuning (Unpaired similarity: 20)

SUF	bn	de	en	id	ko	pt	ru	zh
bn	100	50	43	54	61	54	55	52
de	50	100	78	71	65	71	74	70
en	43	78	100	69	59	68	70	70
id	54	71	69	100	68	71	72	67
ko	61	65	59	68	100	67	67	66
pt	54	71	68	71	67	100	72	67
ru	55	74	70	72	67	72	100	70
zh	52	70	70	67	66	67	70	100

(b) xGQA: Scheduled Unfreezing (Unpaired similarity: 22)

SF	ar	en	es	fr	ru
ar	100	41	48	47	48
en	41	100	48	70	56
es	48	48	100	49	49
fr	47	70	49	100	58
ru	48	56	49	58	100

(c) XVNLI: Standard Finetuning (Unpaired similarity: 17)

SUF	ar	en	es	fr	ru
ar	100	76	83	79	83
en	76	100	79	89	84
es	83	79	100	82	83
fr	79	89	82	100	85
ru	83	84	83	85	100

(d) XVNLI: Scheduled Unfreezing (Unpaired similarity: 62)

Figure 4: Average pairwise CLS-similarity (in percentage points) between the translation-parallel examples of xGQA and XVNLI, compared between scheduled unfreezing (SUF) and standard fine-tuning (SF), evaluated on the last epoch of fine-tuning with UC2. For a baseline of similarity between unpaired examples, we report the average similarity between all examples over all languages (unpaired similarity).

With UC2, the predictions are made from the transformed vector of the sequence start token [CLS]. We thus analyze how similar representations of the [CLS] token are for parallel sentences (same meaning, but in different languages): The more language-agnostic the representations are, the more aligned should the [CLS] token vectors of parallel sentences be.

For this analysis, we leverage the multi-parallel instances of xGQA and XVNLI. We use simple cosine similarity to quantify the similarity of [CLS] vectors of mutual translations. Given that it is possible that a fine-tuning procedure can make inputs appear generally more similar, we also measure "baseline" average similarity between non-parallel sentences (randomly sampled).

Figure 4 displays the results of this analysis on the multi-parallel xGQA and XVNLI test data. We

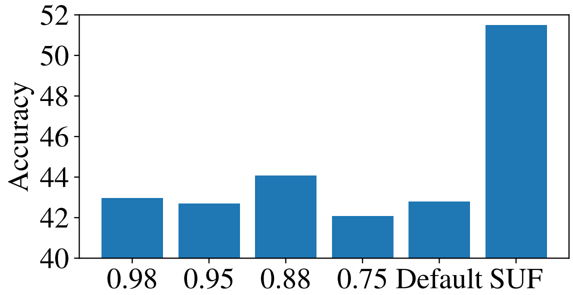


Figure 5: Result of different values for the decay factor d for layer-wise learning rate decay on zero-shot performance for xGQA compared to standard fine-tuning and scheduled unfreezing (SUF). Note that the y-axis starts at 40 to better show performance differences.

make two observations. First, the average similarity with English is highly correlated with the relative zero-shot performance between the languages with a Pearson correlation of over 0.9. This, unsurprisingly, means that there are higher cross-lingual similarities between instances, e.g., for English-German in xGQA or English-French for XVNLI, which also means better transfer results. This confirms the common assumption that good semantic alignment between representations of different languages is key for successful cross-lingual transfer: we show that the same is true for mVLMs. Second, we see that for xGQA, the pairwise similarity between the languages increases substantially more for SUF fine-tuning than for standard fine-tuning (also relatively, compared to the baseline similarity). This suggests that scheduled unfreezing yields more language-agnostic final representations for this task. For XVNLI, where SUF yielded no gains for UC2, the pairwise similarity also increases but so does the baseline similarity, suggesting no improvement in cross-lingual semantic alignment.

5.2 Layer-wise Learning Rate Decay

Our experiments suggest that ZS-XLT, especially with xGQA, profits when the lower layers are trained less. As an alternative to SUF, where a layer is either trained or not (with the same learning rate for all layers), we consider layer-wise learning rate decay. Here, the model is fully trained but we decay the learning rate exponentially between the layers, with a decay factor d, so that parameters of lower layers are trained with much smaller learning rates: For N layers and learning rate l, the actual learning rate $l(i)$ for layer i (counted bottom to top) is: $l(i) = ld^{N-i}$. This means that the top

Setup	en	avg
Standard	62.0	42.8
SUF	**62.8**	**51.5**
CM only	61.9	49.7

Table 2: Results with CCLM on xGQA comparing standard finetuning, scheduled unfreezing (SUF) , and cross-modal layers only (CM only), where we only train the top 6 cross-modal layers and freeze the rest.

layers are trained throughout with the same learning rate as in SUF, but the lower layers, instead of being "flicked-on", after some number of epochs, are instead trained from the start but with a much smaller learning rate. This, in principle, should also limit the overfitting to language-specific knowledge from lower layers.

To evaluate a reasonable range for the decay, we train $\text{CCLM}_{\text{base}}$ on xGQA and choose: $d \in \{0.98, 0.95, 0.88, 0.75\}$ with otherwise the same hyperparameters. As a result, the learning rate of the bottom layer (of 18) is 70% to 0.5% of the learning rate for the top layer.

We present the results in Figure 5. For $d = 0.98$, which decays the least, we see results close to the standard fine-tune setup. For $d = 0.75$, which effectively does not train the lowest layers, performance decreases. We see the best results for $d = 0.88$. While it achieves better results than the standard setup, it underperforms compared to scheduled unfreezing. Looking at per-language results here, we again observe that accuracy for languages like Bengali and Korean, which drop during standard training, are better retained with layer-wise decay.

5.3 Training Top-Layers Only

In Table 2, we test for CCLM, which has 12 XLM-R-initialized text-only layers and 6 cross-modal layers, a setup where we only train the upper 6 cross-modal layers (*CM only* in Table 2). While results are notably better compared to standard fine-tuning for zero-shot transfer, they are slightly worse than with SUF. Allowing the model to adapt the full model, albeit not fully from the start, is important for best performance though results on English are close to standard finetuning.

5.4 SUF in Few-Shot Training

While the focus of this work is on zero-shot cross-lingual transfer, we want to briefly explore if SUF can also further improve results in a few-shot setup. In a few-shot setup, the model is first trained on the large English train split (as in zero-shot) but then also trained on a few dozen to hundred examples in the target language. This can help reduce the performance gap for multiple IGLUE tasks (Bugliarello et al., 2022; Zeng et al., 2022).

Setup	Zero-Shot	Few-Shot
Standard	31.9	44.3
SUF	**41.3**	**46.7**

Table 3: Results for UC2 on xGQA for zero-shot and few-shot when trained with and without SUF on the English train split (*not* for few-shot step).

Following the few-shot setup in IGLUE for xGQA with UC2 (with the maximum 48 shots), we compare a model trained on the English data with and without scheduled unfreezing. During the few-shot training, both setups are trained identically, that is, scheduled unfreezing is not used. As shown in Table 3, SUF is only around 2 points better after few-shot training. While the more language-agnostic representations learned with SUF might be a slightly better starting point for few-shot training, we also see that with a few examples, the model can 'rectify' the performance drop seen during training on English for most languages.

6 Conclusion

Cross-lingual zero-shot allows us to train massively multilingual vision-language models on English task-specific data and then use them for other languages without additional target language training data. Still, there is a large performance gap to English. In this work, we leverage scheduled unfreezing – a finetuning strategy where we initially keep all but the upper model layers frozen and gradually unfreeze the model top-down during training – as a method for reducing the transfer gap.

Experiments with two different models on three downstream vision-language tasks show that scheduled unfreezing can help improve non-English performance; results in visual question answering are especially promising with massive gains in accuracy. Subsequent analysis suggests that scheduled unfreezing can help the zero-shot transfer by forcing the model to learn more language-agnostic features and overfit less on English-specific idiosyncrasies in the training data.

Acknowledgements

This work was in part supported by the Alexander von Humboldt Foundation.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *CoRR*, abs/2308.12966. ArXiv: 2308.12966.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. In *International Conference on Machine Learning*, pages 2370–2392. PMLR.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, A. J. Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023. PaLI-X: On Scaling up a Multilingual Vision and Language Model. *CoRR*, abs/2305.18565. ArXiv: 2305.18565.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *CoRR*, abs/2305.06500. ArXiv: 2305.06500.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavas. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *CoRR*, abs/2307.06930.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with multiq. *CoRR*, abs/2403.03814.

Jeremy Howard and Sebastian Ruder. 2018a. 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.

Jeremy Howard and Sebastian Ruder. 2018b. 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.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Tannon Kew, Florian Schottmann, and Rico Sennrich. 2023. Turning English-centric LLMs Into Polyglots: How Much Multilinguality Is Needed? *CoRR*, abs/2312.12683. ArXiv: 2312.12683.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020*

Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4483–4499.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *CoRR*, abs/2301.12597. ArXiv: 2301.12597.

Zejun Li, Zhihao Fan, Jingjing Chen, Qi Zhang, Xuanjing Huang, and Zhongyu Wei. 2023b. Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5939–5958, Toronto, Canada. Association for Computational Linguistics.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Chen Liu, Jonas Pfeiffer, Anna Korhonen, Ivan Vulić, and Iryna Gurevych. 2023a. Delving deeper into cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2408–2423.

Chen Cecilia Liu, Jonas Pfeiffer, Ivan Vulić, and Iryna Gurevych. 2024. Fun with fisher: Improving generalization of adapter-based cross-lingual transfer with scheduled unfreezing.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021a. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021b. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved Baselines with Visual Instruction Tuning. *CoRR*, abs/2310.03744. ArXiv: 2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual Instruction Tuning. *CoRR*, abs/2304.08485. ArXiv: 2304.08485.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M. Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual Generalization through Multitask Finetuning. *CoRR*, abs/2211.01786. ArXiv: 2211.01786.

Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3p: Learning universal representations via multitask multilingual multimodal pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3977–3986.

Farhad Nooralahzadeh and Rico Sennrich. 2023. Improving the cross-lingual generalisation in visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13419–13427.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xgqa: Cross-lingual visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina

McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *CoRR*, abs/2211.05100. ArXiv: 2211.05100.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual Instruction Tuning With Just a Pinch of Multilinguality. *CoRR*, abs/2401.01854. ArXiv: 2401.01854.

Bin Shan, Yaqian Han, Weichong Yin, Shuohuan Wang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. ERNIE-UniX2: A Unified Cross-lingual Cross-modal Framework for Understanding and Generation. *CoRR*, abs/2211.04861. ArXiv: 2211.04861.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A Corpus for Reasoning About Natural Language Grounded in Photographs. *arXiv:1811.00491 [cs]*. ArXiv: 1811.00491.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2023. X 2-vlm: All-in-one pre-trained model for vision-language tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yan Zeng, Wangchunshu Zhou, Ao Luo, and Xinsong Zhang. 2022. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. *arXiv e-prints*, pages arXiv–2206.

Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165.

A Per-Language Results

We report the per-language results for all our models and tasks.

Zero-Shot	en	ar	es	fr	ru	Ø
UC2	77.1	**56.6**	58.1	68.1	**64.9**	**61.9**
UC2 + SUF	**77.2**	55.6	**58.5**	**69.2**	63.7	61.7
CCLM	**81.2**	60.9	69.6	75.6	68.5	68.6
CCLM + SUF	80.6	**63.6**	**70.9**	**77.6**	**70.4**	**70.6**

Table 4: Accuracy of SUF compared with our baseline on XVNLI on $CCLM_{base}$ and UC2.

Zero-Shot	en	de	bn	id	ko	pt	ru	zh	Ø
UC2	**57.1**	44.4	20.8	30.7	25.3	34.1	35.4	32.8	31.9
UC2 + SUF	**57.1**	**51.6**	**26.5**	**40.5**	**38.6**	**41.2**	**43.8**	**47.0**	**41.3**
CCLM	62.0	57.2	33.7	49.8	29.1	46.4	39.9	43.3	42.8
CCLM + SUF	**62.8**	**59.0**	**49.5**	**52.5**	**42.2**	**51.7**	**48.2**	**57.5**	**51.5**

Table 5: Zero-shot evaluation of scheduled unfreezing on CCLM and UC2.

Zero-Shot	en	de	es	id	ja	ru	tr	zh	Ø
Text Retrieval									
UC2	**36.8**	25.8	16.0	12.8	21.6	16.9	7.3	25.8	18.0
UC2 + SUF	36.4	**26.0**	**17.8**	**16.3**	**23.5**	**19.7**	**8.2**	**29.0**	**20.0**
CCLM	77.7	68.8	66.4	55.3	69.6	64.5	45.6	73.6	63.4
CCLM + SUF	**78.5**	**71.0**	**69.5**	**58.1**	**71.1**	**68.9**	**50.7**	**73.2**	**66.1**
Image Retrieval									
UC2	**43.0**	39.3	15.9	12.7	26.3	19.7	6.4	33.4	20.0
UC2 + SUF	41.8	**30.2**	**18.7**	**15.1**	**28.1**	**22.8**	**8.0**	**33.5**	**22.3**
CCLM	78.0	69.2	68.6	54.8	72.7	64.8	45.7	73.7	64.2
CCLM + SUF	**78.6**	**70.5**	**70.9**	**60.0**	**74.3**	**68.7**	**50.4**	**74.6**	**67.1**

Table 6: Results of SUF compared with our baseline on text and image retrieval (r@1, xFlickrCo) on $CCLM_{base}$ and UC2.