Distilling Multilingual Vision–Language Models: When Smaller Models Stay Multilingual

Sukrit Sriratanawilai$^\heartsuit$**, Jhayahgrit Thongwat**$^\heartsuit$**, Romrawin Chumpu**$^\heartsuit$**,**
Patomporn Payoungkhamdee$^\heartsuit$**, Sarana Nutanong**$^\heartsuit$**, Peerat Limkonchotiwat**$^\spadesuit$
$^\heartsuit$VISTEC, $^\spadesuit$AI Singapore
sukrit.s_s19@vistec.ac.th, peerat@aisingapore.org

Abstract

Vision–language models (VLMs) exhibit uneven performance across languages, a problem that is often exacerbated when the model size is reduced. While Knowledge distillation (KD) demonstrates promising results in transferring knowledge from larger to smaller VLMs, applying KD in multilingualism is an underexplored area. This paper presents a controlled empirical study of KD behavior across five distillation approaches, isolating their effects on cross-lingual representation consistency and downstream performance stability under model compression. We study five distillation formulations across CLIP and SigLIP2, and evaluate them on in-domain retrieval and out-of-domain visual QA. We find that some configurations preserve or even improve multilingual retrieval robustness despite halving model size, but others fail to maintain cross-task stability, exposing design-sensitive trade-offs that aggregate accuracy alone does not reveal.

1 Introduction

Vision–language models (VLMs) have become the dominant paradigm for joint visual–textual representation learning (Radford et al., 2021; Zhai et al., 2023). One prominent approach to achieving performance gains is to utilize a large-scale multilingual corpus (Tschannen et al., 2025). This practice results in a reliance on massive encoder models.

VLM architectures vary widely in scale, from small models such as CLIP (Radford et al., 2021), FILIP (Yao et al., 2021), ALIGN (Jia et al., 2021) to large models like SigLIP (Zhai et al., 2023), SigLIP2 (Tschannen et al., 2025), CoCa (Yu et al., 2022) Notably, a single text encoder can account for more than half of the total model size (e.g., 565M of 881M parameters, or 64%, in SigLIP2-L/16). In contexts where compute is inherently

bounded, models at this scale are not merely inefficient but unusable (Wang et al., 2020; Sun et al., 2020; Chen et al., 2022).

Smaller models present both benefits and challenges. Table 1 reports preliminary results that illustrate this trade-off. Compressing SigLIP2-L/16 yields a substantial inference speedup but also degrades performance. Crucially, we observe degradation in both the image-to-text (I2T) retrieval and downstream CVQA accuracy. In particular, cross-lingual retrieval on the Multi30K dataset decreases by 12.58 points (from 73.43 to 60.85) for I2T and by 4.36 points (from 32.88 to 28.52) for CVQA when the number of parameters is reduced from 881M to 433M using the feature distillation (FD) method proposed by Carlsson et al. (2022).

Params (M)	Inference speed	Recall@1 I2T	CVQA Accuracy
881 (Teacher)	$1\times$	73.43	32.88
593 (Student)	$\sim 3.5\times$	70.90	30.30
450 (Student)	$\sim 6.1\times$	63.38	28.60
433 (Student)	$\sim 9.7\times$	60.85	28.52

Table 1: Effect of compressing SigLIP2-L/16 on inference speed, recall@1 (Multi30K), and accuracy (CVQA). While model size reduction yields a substantial speedup, the performance of downstream tasks decreases as the model size is reduced.

This challenge becomes even more pronounced in multilingual settings. In such environments, KD must also maintain multilingual consistency in cross-modal representations while reducing model size. While KD receives significant attention (Sanh et al., 2020; Wang et al., 2020; Tan et al., 2023; Yang et al., 2024), multilingual consistency is rarely an explicit consideration in existing KD objectives. This consideration is crucial for enabling the efficient deployment of such models in linguistically diverse environments.

In this paper, we tackle this problem through the following two research questions:

- **RQ1:** Across different knowledge distilla-

tion strategies, which ones best preserve multilingual retrieval performance?

- **RQ2:** What impact does the knowledge distillation have on cross-lingual efficiency, latent structure, and retrieval robustness?

In particular, we focus on investigating knowledge distillation in a multilingual vision-language model setting, specifically examining how knowledge can be effectively transferred from vision-language foundation models to multilingual encoder models. To address **RQ1**, we investigate knowledge distillation techniques to improve the performance of multilingual downstream tasks, where we use five techniques covering major KD developments, as well as cross-lingual knowledge transfer solutions. Then, we analyze the performance gap when transferring knowledge from VLM text representation to small multilingual encoders, such as XLM-R$_{Base}$ (Conneau et al., 2019), DistilBERT (Sanh et al., 2020), and MiniLM (Wang et al., 2020). To answer **RQ2**, we analyze the trade-offs involved, identifying when model size reduction preserves multilingual retrieval quality and when it leads to a performance decrease in some languages. We also examine which tasks, such as ranking and clustering, enable smaller models to remain competitive with, or even outperform, larger models.

We evaluate knowledge distillation across two representative teacher models with contrasting multilingual properties: CLIP-ViT-L/14, an English-centric model, and SigLIP2-L/16, a natively multilingual VLM. Our experiments cover both in-domain (image–text retrieval) and out-of-domain (visual question answering) benchmarks, allowing us to examine not only cross-lingual alignment preservation but also generalization beyond the teacher's training distribution. Our findings indicate that performance retention after compression is indeed achievable, but only under the right distillation configurations, with learning objectives playing a particularly critical role. Moreover, the effect is strongly task-dependent: while some settings suffer degradation, tasks such as multilingual reranking and clustering prove more resilient, with smaller models in some cases performing competitively with larger ones.

Our contributions are as follows:

- We present the first systematic comparative study of knowledge distillation strategies in multilingual vision–language models, char-

acterizing how different design choices influence cross-lingual alignment and performance under compression.

- We analyze trade-offs, showing when size reduction preserves multilingual retrieval quality, when some languages perform poorly, and which tasks (e.g., ranking, clustering) allow smaller models to remain competitive with or outperform larger models.

2 Related Work

2.1 Multilingual Vision-language Model Training

Recently, many works (Radford et al., 2021; Jia et al., 2021; Li et al., 2023) have researched multimodality (vision-text encoder) by aggregating ViT (Dosovitskiy et al., 2021) and text encoder (Devlin et al., 2019) and training the model using vision-text datasets (i.e., CC-12M (Changpinyo et al., 2021) or LAION (Schuhmann et al., 2022)). The training objective of this kind of work is to maximize the similarity of text and image pairs, while minimizing the similarity of irrelevant pairs using contrastive learning (Radford et al., 2021). This technique has been proven to be a robust training technique to achieve a strong vision-text encoder model on retrieval (Iscen et al., 2024) or VQA (Kant et al., 2021; Parelli et al., 2023). However, these previous works only experimented in English, while the multilingual capability might not have been well established.

Researchers extend the CLIP method to support multilingualism by aligning CLIP's representations with those of multilingual texts in the same embedding space. Multilingual CLIP (Carlsson et al., 2022) uses knowledge distillation from CLIP text's encoder as teacher and multilingual encoder as student by using text pairs that are translated by machine translation as training data. mCLIP (Chen et al., 2023) aligns English text representation between CLIP's encoder and multilingual encoder with Triangle cross-modal Knowledge Distillation loss. Recently, SIGLIP2 (Tschannen et al., 2025) has been proposed as a multilingual and multimodal foundation model trained from scratch on multilingual text-image pairs using sigmoid loss and self-distillation methods, unlike CLIP, which employs only InfoNCE loss. However, these models rely on a large text encoder, where the size of the text encoder accounts for 64% of the total parameters.

We require an exploration of how to integrate the small model into these techniques.

2.2 Knowledge Distillation

Knowledge Distillation (KD) (Hinton, 2015) is a cross-architecture knowledge transfer technique that transfers knowledge from a teacher model to a student model. KD methods achieve this by guiding the student's learning process to align its representations with those of the teacher, using a training objective that minimizes the discrepancy between these two models. A common technique in the KD manner is to train the student model to maximize the similarity between the teacher and student probability distributions. Feature-based distillation (FD) (Romero et al., 2015) aligns the teacher and student models' representations by minimizing the mean squared error between the embeddings of the teacher and student. CRD (Tian et al., 2019) using the contrastive objective as a learning function to align teacher and student representations. RKD (Park et al., 2019; Yang et al., 2022) aligns the teacher and student feature representations with the transfer mutual relation of the feature from the teacher to the student model. DualL2 (Reimers and Gurevych, 2020) uses FD to minimize the mean squared error between the teacher's English sentence representation and the student's parallel language sentence.

To enhance the performance of knowledge distillation (KD) methods, researchers have proposed various techniques in the KD pipeline. For example, adding augmentation techniques to generalize students' representation (Jiao et al., 2020), improve domain-specific KD method (Weng et al., 2024), or use in low-resources language KD (Tan et al., 2023). Moreover, researchers employ self-distillation with a momentum encoder (Li et al., 2021) and add an instance queue to increase the diversity of negative samples for the KD loss (Fang et al., 2021; Limkonchotiwat et al., 2022). While knowledge distillation has proven effective for improving both cross-lingual and small models, the most effective approach for the multilingual vision-text encoder remains an open question.

3 Methodology

3.1 Problem Formulation

To decrease the model's parameters, we apply the concept of knowledge distillation to transfer the knowledge from a larger model to a smaller one.

In particular, we minimize the discrepancy between large and small models, where the input can be more than one language for the student model. Let $\mathcal{D} = \{(x_{i,\text{e}}, x_{i,\text{m}})\}_{i=1}^{N}$ denote a dataset consisting of N paired English and multilingual text samples. The vector representations produced by the student model are obtained via the embedding function $f(\cdot; \theta_{\text{S}})$, yielding $z_{i,\text{e}}^{\text{S}} = f(x_{i,\text{e}}; \theta_{\text{S}})$ for English inputs and $z_{i,\text{m}}^{\text{S}} = f(x_{i,\text{m}}; \theta_{\text{S}})$ for multilingual inputs, where θ_{S} denotes the parameters of the student model. Similarly, the English representation from the teacher model, parameterized by θ_{T}, is given by $z_{i,\text{e}}^{\text{T}} = f(x_{i,\text{e}}; \theta_{\text{T}})$.

To facilitate knowledge transfer from the teacher to the student, the discrepancy between their representations is minimized using multiple objective functions, formally expressed as

$$\min_{\theta_{\text{S}}} \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}\left(z_{i,\text{m}}^{\text{S}}, z_{i,\text{e}}^{\text{S}}, z_{i,\text{e}}^{\text{T}}; \theta_{\text{S}}\right),$$

where B denotes the batch size. In all experiments, the optimization objective is to minimize the loss with respect to the student model parameters θ_{S}, while keeping the teacher model parameters θ_{T} fixed.

3.2 Knowledge Distillation Loss

As shown in Figure 1, in this study, we examine alternative loss functions for multilingual vision–language embedding distillation. Our investigation centers on various feature distillation approaches, contrastive learning, and distributional replication loss, with their effectiveness evaluated across multilingual benchmarks.

3.2.1 Feature Distillation (FD)

A straightforward approach to transferring knowledge from the teacher model to the student model is to anchor the English representations generated by the teacher and minimize their discrepancy with the student's representations for the corresponding multilingual inputs.

The discrepancy between teacher and student representations is minimized using the Mean Squared Error loss, defined as:

$$\mathcal{L}_{\text{FD}} = \frac{1}{B} \sum_{i=1}^{B} \left\| z_{i,\text{m}}^{\text{S}} - z_{i,\text{e}}^{\text{T}} \right\|_2^2 \qquad (1)$$

3.2.2 English-Control Distillation (ED)

This approach extends Feature Distillation by incorporating the alignment of the student's English
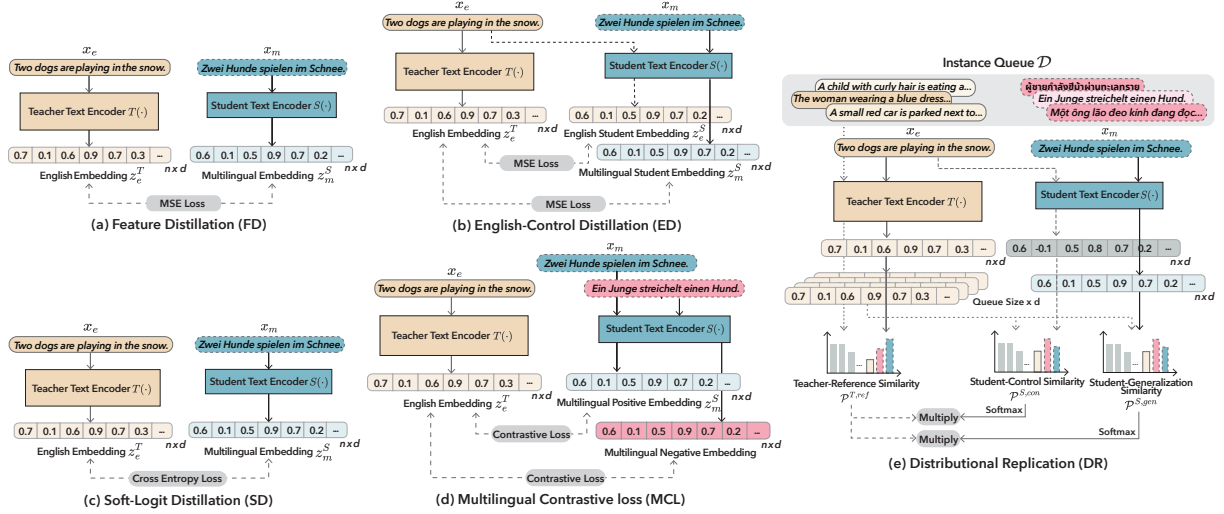
Figure 1: Illustration of variation multilingual vision-language embedding distillation in this paper

text representations in addition to its multilingual representations. The inclusion of English representation alignment helps prevent representation collapse, which occurs when different language representations are drawn toward the same anchor, namely the teacher's English representation. Derived from feature-based distillation, the training objective is formalized as:

$$\mathcal{L}_{\text{ED}} = \frac{1}{B} \sum_{i=1}^{B} \left(\left\| z_{i,\text{m}}^{\text{S}} - z_{i,\text{e}}^{\text{T}} \right\|_2^2 + \left\| z_{i,\text{e}}^{\text{S}} - z_{i,\text{e}}^{\text{T}} \right\|_2^2 \right) \tag{2}$$

3.2.3 Soft-Logit Distillation (SD)

We also employ the Cross-Entropy loss, which is particularly suitable when the knowledge to be distilled is represented as probability distributions. This approach assumes that the teacher and student models produce vector representations that can be softened into categorical distributions:

$$p_{i,\text{e}}^{\text{T}} = \text{softmax}(z_{i,\text{e}}^{\text{T}}), \quad p_{i,\text{m}}^{\text{S}} = \text{softmax}(z_{i,\text{m}}^{\text{S}})$$

The soft-logit distillation loss measures the dissimilarity between the teacher and student probability distributions and is defined as:

$$\mathcal{L}_{\text{SD}} = -\frac{1}{B} \sum_{i=1}^{B} p_{i,\text{e}}^{\text{T}} \log p_{i,\text{m}}^{\text{S}} \tag{3}$$

3.2.4 Multilingual Contrastive Learning (MCL)

To align student and teacher representations, we employ contrastive learning (CL), which optimizes the objective by maximizing similarity between positive teacher–student pairs derived from

the same English input, while contrasting them against other in-batch negative pairs. The objective is formulated as:

$$\mathcal{L}_{i,\text{e}}^{\text{MCL}} = -\log \frac{\exp(\text{sim}(z_{i,\text{e}}^{\text{S}}, z_{i,\text{e}}^{\text{T}})/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(z_{i,\text{e}}^{\text{S}}, z_{j,\text{e}}^{\text{T}})/\tau)} \tag{4}$$

where τ denotes the temperature parameter and cosine similarity is adopted as the similarity function.

In addition, to enable knowledge transfer from teacher to student, the student's multilingual representation obtained from the same English input processed by the teacher is used to compute a parallel contrastive objective:

$$\mathcal{L}_{i,\text{m}}^{\text{MCL}} = -\log \frac{\exp(\text{sim}(z_{i,\text{m}}^{\text{S}}, z_{i,\text{e}}^{\text{T}})/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(z_{i,\text{m}}^{\text{S}}, z_{j,\text{e}}^{\text{T}})/\tau)} \tag{5}$$

The overall Multilingual Contrastive Learning (MCL) loss combines both English-based and multilingual objectives, expressed as:

$$\mathcal{L}_{\text{MCL}} = \frac{1}{B} \sum_{i=1}^{B} (\mathcal{L}_{i,\text{e}}^{\text{MCL}} + \mathcal{L}_{i,\text{m}}^{\text{MCL}})/2 \tag{6}$$

3.2.5 Distributional Replication (DR)

Distributional Replication (DR) quantifies the divergence between teacher and student outputs by constructing similarity-based probability distributions. These distributions are generated from a FIFO queue of negative samples, $\mathbf{Q} = [q_1, ..., q_K]$, which is continuously updated with the teacher's in-batch English representations, $[z_{1,\text{e}}^{\text{T}}, ..., z_{B,\text{e}}^{\text{T}}]$. The generic probability distribution is defined as:

$$\mathcal{P}_{ik}(\boldsymbol{z}, \boldsymbol{Q}, \tau) = \frac{\exp(\text{sim}(z_i, q_k)/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(z_i, q_j)/\tau)} \tag{7}$$

Within this framework, DR specifies three distinct distributions, each serving a complementary role: (i) Teacher-Reference distribution, computed from z_e^{T}, which acts as the teacher-provided reference:

$$\mathcal{P}_{ik}^{\text{T,ref}} = \mathcal{P}_{ik}(z_e^{\text{T}}, \boldsymbol{Q}, \tau^{\text{T}}) \qquad (8)$$

(ii) Student-Control distribution, computed from z_e^{S}, which constrains student representations to remain aligned with teacher knowledge:

$$\mathcal{P}_{ik}^{\text{S,con}} = \mathcal{P}_{ik}(z_e^{\text{S}}, \boldsymbol{Q}, \tau^{\text{S}}) \qquad (9)$$

(iii) Student-Generalize distribution, computed from z_m^{S}, which facilitates broader generalization in the student's multilingual space:

$$\mathcal{P}_{ik}^{\text{S,gen}} = \mathcal{P}_{ik}(z_m^{\text{S}}, \boldsymbol{Q}, \tau^{\text{S}}) \qquad (10)$$

The DR objective consists of two complementary components. The control objective enforces consistency between student-control and teacher-reference distributions:

$$\mathcal{L}_i^{\text{con}} = -\sum_{k=1}^{K} \mathcal{P}_{ik}^{\text{T,ref}} \log \mathcal{P}_{ik}^{\text{S,con}} \qquad (11)$$

while the generalization objective extends this consistency to the student-generalize distribution:

$$\mathcal{L}_i^{\text{gen}} = -\sum_{k=1}^{K} \mathcal{P}_{ik}^{\text{T,ref}} \log \mathcal{P}_{ik}^{\text{S,gen}} \qquad (12)$$

Finally, the overall DR loss is expressed as the average of the two objectives:

$$\mathcal{L}_{\text{DR}} = \frac{1}{B} \sum_{i=1}^{B} \left(\mathcal{L}_i^{\text{con}} + \mathcal{L}_i^{\text{gen}} \right) / 2 \qquad (13)$$

3.3 Multi-Objective Training

As we discussed the benefits and strengths of each training objective, we found that each loss has a trade-off, and there is no universal solution to the vision-text representation problem. Alternatively, (Yang et al., 2024; Limkonchotiwat et al., 2024) demonstrate the possibility of combining each training loss as a multi-task training objective. Therefore, we summarize all training objectives with a joint knowledge distillation objective in this section:

$$\mathcal{L} = \sum_{i=0}^{n} \lambda_i \mathcal{L}_i \qquad (14)$$

Where $\mathcal{L}_i$ represents the distillation objective that we mentioned previously, and λ_i are their weight for each objective.

4 Experimental Setup

4.1 Training Dataset

Following previous works (Carlsson et al., 2022; Chen et al., 2023; Zhai et al., 2023), we utilize a common training dataset, Imagecaptioning7M, which comprises 7M multilingual-English text pairs. We used the translated version from Carlsson et al. (2022). This dataset was derived from sources like Google Conceptual Caption (GCC) (Sharma et al., 2018), MSCOCO (Lin et al., 2014), and VizWiz (Bigham et al., 2010). For the Validation dataset, we use a validation set of Multi30k (Elliott et al., 2016), a multilingual version of Flickr30k (Plummer et al., 2015).

4.2 Models

We utilize CLIP (Radford et al., 2021), a monolingual foundation model, and SigLIP2 (Tschannen et al., 2025), a multilingual foundation model. In particular, we select CLIP-ViT-L/14 and SigLIP2-L16 as teacher models for our experiments, and we select XLM-R$_{\text{Base}}$ (Conneau et al., 2019), MiniLM (Wang et al., 2020), and DistilBERT (Sanh et al., 2020) as student models to mimic the teacher text's representation. We describe the hyper-parameter settings in Appendix A.

4.3 Evaluation Benchmark

Similar to previous works' setting (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025), we evaluate our student models on seven benchmarks that cover text-image retrieval and Visual Question Answering (VQA) tasks. For retrieval downstream tasks, we use Multi-30k (Elliott et al., 2016), MSCOCO (Lin et al., 2014), WIT (Srinivasan et al., 2021), xFlickr (Bugliarello et al., 2022), and XM3600 (Thapliyal et al., 2022). For VQA, we utilize CVQA (Romero et al., 2024), which comprises user-submitted photos and questions in 31 languages, and ALM-Bench (Vayani et al., 2024), which offers domain-specific cultural questions in 100 languages.

4.4 Evaluation metrics

We employ Recall@k ($R@k$) for Text-to-Image (T2I) and Image-to-Text retrieval (I2T) tasks. Our primary metric is $R@1$, which measures top-1 accuracy, and we also report results for $R@5$ and $R@10$ in Appendix B. For VQA, we formulate this task as a similarity matching problem similar

| Methods | Retrieval | | | | | | | | VQA | | |
| | Multi30k | | COCO | | WIT | xFlickr | XM3600 | | CVQA | ALM-Bench | |
	I2T	T2I	I2T	T2I	I2T	I2T	I2T	AVG.	Acc	Acc	AVG.
T: SigLIP2-L/16	71.37	71.23	45.82	20.37	42.39	53.64	53.75	51.22	32.88	39.53	36.21
S: XLM-R$_{Base}$											
+FD	69.27	73.20	**36.12**	25.89	21.99	62.24	48.45	48.17	30.30	36.70	33.50
+ED	63.97	74.67	32.58	**29.20**	30.26	65.24	52.65	49.80	**31.01**	**40.06**	**35.54**
+SD	58.80	68.37	27.12	24.14	20.53	57.59	44.62	43.02	29.49	36.73	33.11
+MCL	54.37	63.27	26.92	24.63	19.68	54.24	42.36	40.78	26.87	35.81	31.34
+DR	**70.57**	76.03	30.34	27.39	**33.40**	66.66	54.13	**51.22**	28.65	36.48	32.57
+DR+ED	**71.03**	76.17	30.44	27.17	33.41	66.48	54.22	51.27	28.76	36.13	32.45
+DR+FD	70.47	76.23	31.08	27.30	33.64	66.51	54.23	**51.35**	28.67	36.46	32.57
+DR+ED+FD	70.90	**76.53**	30.14	26.51	33.20	66.22	**54.26**	51.11	**28.68**	**37.53**	**33.11**

Table 2: This table presents our qualitative evaluation of two metrics: (1) retrieval Recall@1 (R@1) scores on Multi30K, COCO, WIT, xFlickr, and XM3600; and (2) multiple-choice visual question answering accuracies on CVQA and ALM-Bench. The results are from the XLM-R$_{Base}$ student model trained with knowledge distillation from the SigLIP2-L/16 teacher model.

to Romero et al. (2024). In particular, to determine the model answer, we concatenate the question with candidate answers, compute the cosine similarity between the combined text and image, and select the most similar choice as the answer. Then, we use accuracy as the main metric of VQA benchmarks.

5 Experimental Results

In this section, we propose studies to explore the performance of small models using various proposed KD methods, aiming to answer **RQ1**. In Section 5.1, we propose an empirical study of the effectiveness and generalization of using various KD methods on the baseline model, XLM-R$_{Base}$. Section 5.2, we study the robustness of the optimal KD approaches through various small models. Section 5.3, we study the design choices of our KD method, specifically the language anchor for KD and the representation of images versus text as the anchor.

5.1 Main Results

KD results The results of our KD methods are shown in Table 2. We can see that the student model with the DR method outperformed other individual KD methods. Although the ED method performed less effectively than the DR method on retrieval tasks, it significantly outperformed all other methods on the VQA task. For example, the ED method outperforms the DR method by 2.97 points in the VQA benchmarks. This emphasizes that there is no universal method; the DR method is suitable for in-domain downstream tasks, such as the retrieval task, whereas ED is more generalized than DR in comparison to out-of-domain

VQA tasks.

Compare with the teacher's performance When we compare the performance of the best performing students and the teacher model, we found that *the student can perform similarly to the teacher model in the average score.* As shown in Table 2, the DR method achieves 51.22 points on the retrieval benchmarks, matching the performance of the teacher model. We observe a reasonable improvement on all T2I experiments, including Multi30k and COCO. These findings demonstrate that KD can enable student models to mimic teacher behavior and give better performance than the teacher in some downstream tasks.

Combining multiple KD training objectives We also conduct an experiment using multi-KD training objectives in our study by combining the most effective training objectives in Table 2. The experimental results demonstrate that combining DR and FD yields a better improvement for the retrieval task, improving from 51.22 points with DR to 51.35 points with DR+FD. However, we observe a performance penalty in the VQA task from 35.54 (ED) to 33.11 points (DR+ED+FD). These findings emphasize the importance of the training objective, which is designed for the retrieval task, rather than the VQA task. This suggests the need to develop a new training objective that effectively addresses both retrieval and VQA tasks. Note that we demonstrate the full results of each language in Appendix E.

5.2 Model Variants

To assess the robustness of the KD techniques, we experiment on the same benchmarks using various teacher and student models. In particular, we

Methods	Retrieval (I2T)												VQA					
	Multi30k		COCO		WIT		xFlickr		XM3600		AVG		CVQA		ALM-Bench		AVG	
	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul
SigLIP2-L/16	79.60	71.37	70.48	45.82	70.70	42.39	74.25	53.64	65.72	53.75	72.15	53.39	31.33	32.88	35.88	39.53	33.61	36.21
T: SigLIP2-L/16 / S: XLM-R_Base (Parameters from 881M to 594M):Text Encoder from 565M to 278M																		
+FD	75.80	69.27	**63.04**	**36.12**	35.00	21.99	71.95	62.24	49.76	48.45	59.11	47.61	**31.72**	30.30	35.54	**36.70**	**33.63**	**33.50**
+DR	**76.50**	**70.57**	63.02	30.34	**51.30**	33.40	**75.35**	**66.66**	57.38	54.13	**64.71**	51.02	31.12	28.65	35.42	36.48	33.27	32.57
+DR+FD	75.90	70.47	62.94	31.08	50.70	**33.64**	74.90	66.51	**57.50**	**54.23**	64.39	**51.19**	30.92	28.68	**35.84**	36.47	33.38	32.58
T: SigLIP2-L/16 / S: DistillBert (Parameters from 881M to 450M):Text Encoder from 565M to 134M																		
+FD	69.60	61.30	59.24	**19.48**	32.30	18.03	68.60	56.41	46.74	38.20	55.30	38.68	30.39	**28.60**	34.37	**37.71**	32.38	**33.16**
+DR	74.20	66.27	61.32	13.64	45.00	26.11	74.10	**62.66**	56.18	**45.30**	62.16	42.80	**30.79**	28.52	**36.54**	35.22	**33.67**	31.87
+DR+FD	**75.00**	**66.63**	**61.58**	13.72	**45.90**	**26.76**	**74.95**	62.53	**56.72**	45.27	**62.83**	**42.98**	30.49	27.54	34.60	35.34	32.55	31.44
T: SigLIP2-L/16 / S: MiniLM (Parameters from 881M to 433M):Text Encoder from 565M to 117M																		
+FD	66.80	58.87	52.68	**29.10**	17.30	11.27	63.75	50.08	40.22	37.50	48.15	37.36	27.80	**28.52**	30.90	**35.81**	29.35	**32.17**
+DR	73.10	65.07	58.92	21.90	40.30	22.84	73.80	61.16	**55.79**	**46.73**	60.38	**43.54**	29.42	28.39	**33.60**	34.28	**31.51**	31.34
+DR+FD	**73.60**	**65.43**	**59.60**	20.76	**41.60**	**23.08**	**74.35**	**61.56**	55.58	46.68	**60.95**	43.50	**29.85**	28.12	32.87	34.28	31.36	31.20
CLIP-ViT-L/14	68.80	-	56.32	-	69.50	-	56.00	-	42.47	-	58.62	-	36.72	-	43.69	-	40.21	-
T: CLIP-ViT-L/14 / S: XLM-R_Base (Parameters from 427M to 581M):Text Encoder from 123M to 278M																		
+FD	**67.00**	**62.13**	**53.40**	**33.18**	32.00	22.06	56.80	47.68	37.50	36.39	49.34	40.29	**29.28**	**29.04**	33.87	36.04	**31.58**	**32.54**
+DR	63.30	59.30	49.04	25.02	**54.60**	34.27	**64.40**	53.85	**43.72**	42.07	**55.01**	42.90	28.74	28.13	33.30	36.34	31.02	32.24
+DR+FD	62.70	59.30	49.24	25.94	52.40	**34.36**	63.10	**54.01**	43.61	**42.15**	54.21	**43.15**	28.90	27.92	**34.17**	**36.70**	31.54	32.31
T: CLIP-ViT-L/14 / S: DistillBert (Parameters from 427M to 437M):Text Encoder from 123M to 134M																		
+FD	**63.50**	57.20	**50.94**	**20.04**	31.20	18.34	52.45	41.93	35.04	28.93	46.63	33.29	**29.14**	27.75	32.78	**35.54**	**30.96**	**31.65**
+DR	63.20	55.43	48.58	12.84	**49.10**	29.52	62.80	**51.84**	**43.08**	35.94	**53.35**	37.11	27.37	27.33	**32.85**	34.92	30.11	31.13
+DR+FD	62.20	**55.90**	47.88	12.28	48.80	**30.57**	**62.90**	51.21	42.83	**35.96**	52.92	**37.18**	27.83	27.18	32.82	35.24	30.33	31.21
T: CLIP-ViT-L/14 / S: MiniLM (Parameters from 427M to 420M):Text Encoder from 123M to 117M																		
+FD	58.40	54.50	**47.64**	**28.58**	17.80	11.37	46.75	37.00	30.07	28.05	40.13	31.90	27.49	**28.44**	30.11	**34.86**	28.80	**31.65**
+DR	**62.50**	**55.20**	46.00	21.30	**40.10**	25.18	62.00	50.23	42.53	37.50	**50.63**	**37.88**	**28.21**	26.74	31.14	34.24	**29.68**	30.49
+DR+FD	62.00	53.97	46.22	21.28	39.40	**25.51**	**62.40**	**50.91**	**42.63**	**37.52**	50.53	37.84	27.78	27.36	**31.39**	34.63	29.59	31.00

Table 3: A comparison of knowledge distillation performance for various teacher-student models trained on the ImageCaptioning7M dataset and validated on the Multi30k dataset.

select three KD techniques: (i) FD as a strong baseline, (ii) DR as the most effective individual approach, and (iii) the optimal multi-objective configuration, DR combined with FD. We vary the teacher–student configurations by employing SigLIP2-L/16 and CLIP-ViT-L/14 as teacher models, while considering XLM-R_Base, DistilBERT, and MiniLM as student models for comparison.

Comparing with SigLIP2 As shown in Table 3, we observe that when we decrease the text encoder parameters from 565M (SigLIP2-L/16) to 278M, the multilingual performance decreases from 53.39 to 51.19 points on the retrieval benchmarks, while the performance of English decreases by 7.44 points. However, performance decreases when the number of parameters is reduced; for example, we observed a 16.03-point decrease in FD when using MiniLM. Although we can mitigate this problem with our KD techniques (DR and DR+FD), a gap still remains between the teacher and student models for the retrieval benchmarks. In contrast, we found that only a 2.71-point difference on the VQA benchmarks. Moreover, we obtain a significantly faster inference speed, which is preferable for real-world applications. This can be a trade-off for efficiency vs. robustness for the retrieval task.

Cross-lingual transfer capability Interestingly, we found that when we use CLIP-ViT-L/14 as the teacher model (which only supports English), we can create a student model that supports multiple languages. This is because our learning techniques did not rely on multilingual representation, but instead used only a monolingual representation, enabling the student model to learn any languages supported by the training dataset. The experimental results demonstrate a comparable result between CLIP-ViT-L/14 and students on the English results for the retrieval benchmark.

Multi-training objective is essential When focusing on the multi-objective learning results, it provides an improvement in the setting of smaller student models and in-domain downstream tasks. From Table 3, multi-objective models mostly maintain English performance over single-objective approaches, specifically across all SigLIP2-DistillBert retrieval benchmarks and in four out of five SigLIP2-MiniLM retrieval benchmarks. For CLIP-ViT-L/14 as the teacher model, the multilingual results of the multi-objective approach outperform three of five in retrieval benchmarks (WIT, xFlickr, and XM3600). These results demonstrate that in a small student model, multi-objective learning plays a crucial role in enhancing

Methods	Retrieval (I2T)												VQA					
	Multi30k		COCO		WIT		xFlickr		XM3600		AVG		CVQA		ALM-Bench		AVG	
	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul
Method: DR / T: SigLIP2-L/16 / S: XLM-R$_{Base}$																		
English	**76.50**	**70.57**	63.02	30.34	**51.30**	**33.40**	**75.35**	**66.66**	57.38	54.13	64.71	51.02	**31.12**	28.65	**35.42**	36.48	33.27	**32.57**
German	71.40	70.17	**64.54**	**37.26**	45.70	30.20	71.10	64.19	52.04	52.43	60.96	50.85	30.87	28.58	34.86	36.02	32.87	32.30
China	53.90	47.30	48.12	25.70	25.90	17.35	49.45	40.74	36.76	36.01	42.83	33.42	30.55	**28.96**	34.15	35.11	32.35	32.04
Method: DR+FD / T: SigLIP2-L/16 / S: XLM-R$_{Base}$																		
English	**75.90**	**70.47**	62.94	31.08	**50.70**	**33.64**	**74.90**	**66.51**	57.50	54.23	64.39	51.19	30.92	28.68	**35.84**	36.47	**33.38**	32.58
German	70.60	70.13	**64.20**	**36.06**	45.70	29.59	71.15	63.54	52.40	52.36	60.81	50.37	30.78	**28.96**	34.24	**36.69**	32.51	**32.83**
China	53.50	47.90	48.12	25.58	25.70	17.21	49.55	41.77	37.12	35.95	42.80	33.68	29.90	28.18	35.26	35.79	32.58	31.99

Table 4: A comparison of knowledge distillation performance in SigLIP2-L/16 as teacher and XLM-R$_{Base}$ as student when changing language anchor.

knowledge transfer from teacher to student.

5.3 Ablation study

To confirm the knowledge distillation setting in our work, we conduct ablation studies to observe the performance improvement of each component. We observe that a major component of our KD techniques is the representation of the anchor, while we use English text as the representation for the student model to mimic. Emergent questions raised are: (i) Can we use images instead of text representations? (ii) Is the KD framework generalized to other languages as the anchor?

5.3.1 Non-english language anchor

Setup While a multilingual teacher, SigLIP2-L/16, is able to encode non-English languages, we translated the training dataset into non-English languages to study the knowledge transfer performance when it comes from non-English texts. We select two non-English languages, German, which represents the Indo-European language family, and Chinese, which represents the Sino-Tibetan language family, as training anchors. Then, we translated Imagecaptioning7M to selected languages with Qwen3-4B-Instruct (Yang et al., 2025).

Results As shown in Table 4, using non-English languages leads to performance drops compared to English. The German-trained model outperforms the Chinese-trained model by 17.42 points on the retrieval benchmark, using SigLIP2-L/16 as the teacher, although both remain below the English-trained model. We attribute this to linguistic proximity (German), which is related to English, preserving more knowledge during translation, while Chinese's distinct structure causes greater information loss, as seen in Figure 2. Interestingly, VQA results show only slight declines from En-

glish, with mixed outcomes between German- and Chinese-trained models. Notably, in DR+FD, the German-trained model even surpasses the English model on CVQA and ALM-Bench multilingual tasks, indicating that non-English models can achieve VQA performance comparable to that of the English model and the teacher model.

5.3.2 Using image representation as anchor

Setup Since we use the image-text encoder as the teacher model, we raise the question of whether we can replace a text with an image sample from ImageCaptioning7M to improve performance. From the datasets, we are unable to collect all images in ImageCaptioning7M from the web source, so we can collect approximately 55% of the image data. Therefore, we will compare the text anchor and image anchor with the same total number of training data for a fair comparison.

Results The results in Table 5 show that the student model using an image anchor performs worse than the model using a text anchor in the retrieval benchmarks. We observed a 14.79 point gap between text and image performance using DR, where the decreasing trend is also similar for DR+FD. We hypothesize that the significant performance drop in retrieval tasks is due to using the image representation as an anchor. While the text anchor encodes semantic and grammatical information directly related to the original text, the image anchor provides more ambiguous reference information because its representations encode visual features that can correspond to multiple descriptions. Therefore, the model distilled with text anchors can accurately retrieve the texts or images corresponding to a given pair. However, in the VQA benchmark, the performance of the image-anchor and text-anchor students is comparable. This might be because VQA is the out-of-

| Anchor | Retrieval (I2T) | | | | | | | | | | | | VQA | | | | | |
| | Multi30k | | COCO | | WIT | | xFlickr | | XM3600 | | AVG | | CVQA | | ALM-Bench | | AVG | |
	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul	En	Mul
Method: DR / T: SigLIP2-L/16 / S: DistillBert																		
Text	**73.80**	**63.80**	**59.04**	**14.50**	**49.10**	25.99	**75.00**	60.64	**57.65**	45.80	62.92	42.15	31.12	28.02	35.77	**36.31**	33.46	**32.17**
Image	49.90	38.40	44.36	8.30	30.90	17.25	57.60	42.57	42.53	30.28	45.06	27.36	**31.65**	28.06	37.07	35.04	**34.36**	31.55
Method: DR+FD / T: SigLIP2-L/16 / S: DistillBert																		
Text	74.50	63.90	59.26	13.40	48.70	26.05	75.00	61.59	57.47	45.84	62.99	42.17	30.35	27.67	36.52	35.65	33.44	31.66
Image	48.90	37.27	43.90	8.34	30.70	17.74	58.45	43.74	42.36	30.29	44.86	27.48	30.93	29.16	36.71	35.84	33.82	32.50

Table 5: A comparison of knowledge distillation performance in SigLIP2-L/16 as teacher and DistillBert as student when using images as anchor.

Figure 2: Result of recall@1 on the XM3600 dataset

domain task; using text or image representations cannot mitigate this problem.

6 Analysis

To better understand the results in downstream performance, we provide analyses centered around **RQ2**, as follows:

- Section 6.1, we investigate teacher-student efficiency along with various KD techniques across various languages.
- Section 6.2, we study the language distribution in latent representation.
- Section 6.3, we present the analysis of ranking performance in a retrieval dataset.

6.1 Improvement Across Languages

As shown in Table 2, we observe cases where the student model outperforms the teacher model. This raises the question of why the student model, which has fewer parameters than the teacher model, can outperform it. As shown in Figure 2, we found that the teacher model performs poorly in certain languages. For example, the performance of Japanese and Chinese on xFlickr is lower than that of other languages, resulting in a reasonable improvement for the student. Additionally, the teacher model's performance is poor on Bengali, Telugu, and Swahili; however, these languages are included in our training data, which improves the performance of the student model. This emphasizes the importance of our KD approach on small models, which, although some languages are not well-aligned, can be improved using the available languages in the training data.

6.2 Representation Analysis

To further understand the language distribution in the latent space of a student model, we visualize its embedding using t-SNE on the Multi30k dataset, based on XLM-R$_{Base}$ as a student model, with the most effective approach being the DR. Assuming that *the ideal representation would be distributed well across languages*, quantified by the purity score (Zhao and Karypis, 2001), which indicates the discrepancy of its clustering performance (lower is better).

As shown in Figure 3a, although the retrieval performance of the teacher model is higher than the student model (Figure 3c), when we plot the sample using t-SNE, we observe that the clustering result of the teacher model is poorer than the student model. We can see that the purity score of the student model is lower than that of the teacher model (0.479 vs. 0.321 points). Although we started from a score of 0.942 (Figure 3b), we can enhance it to outperform the teacher model. This emphasizes that the KD method, which focuses on improving multilingual consistency (i.e., all languages exhibit the same distribution), can yield a significant improvement in clustering results. We provide results from other KD methods in Appendix C.
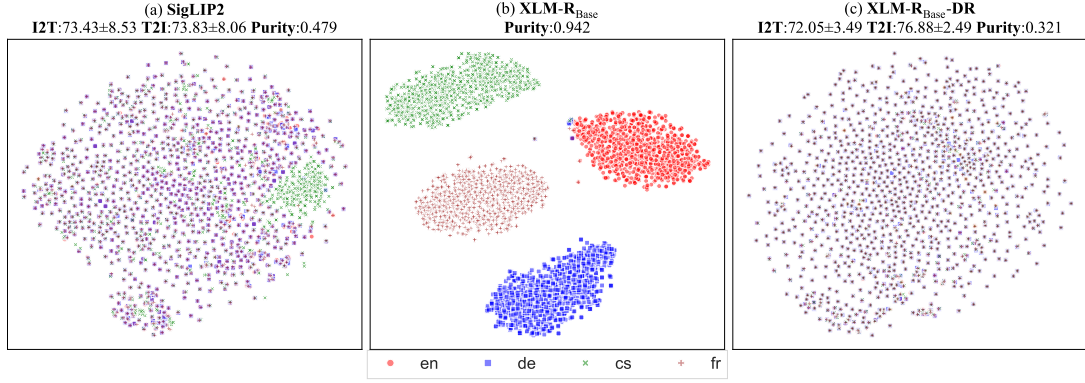
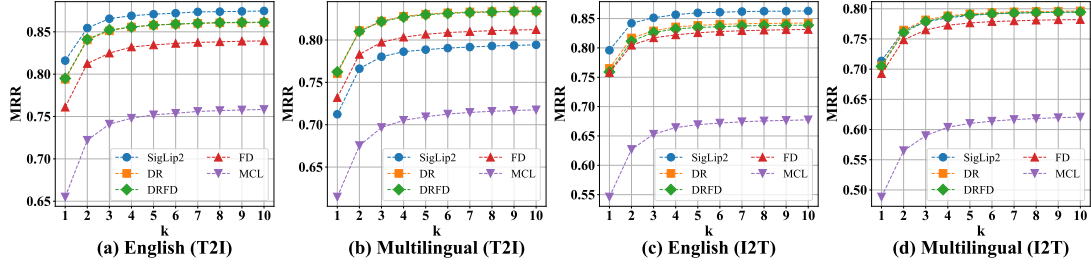Figure 3: Embedding distribution in Multi30k dataset.

Figure 4: Result of MRR@K from Multi30k dataset.

6.3 Ranking Robustness

To empirically examine the robustness and consistency of these properties in downstream tasks, we analyze them via a ranking-based evaluation. As illustrated in Figure 4, the Mean Reciprocal Rank (MRR) metric is employed to quantify the performance of student models across multiple candidate retrieval tasks. The MRR curve shows that the DR-distilled model achieves higher MRR values than other student models across all k. Although the retrieval performance of the DR-distilled student model does not surpass that of the teacher model on the English results, this student model outperforms the SigLIP2-L/16 on the multilingual text-to-image retrieval task and yields comparable retrieval performance on the multilingual image-to-text task. This improvement can be attributed to the use of text-anchor embeddings during distillation, which enhances the text encoder's capability in multilingual settings. Similar to the previous experiment, which utilized a pre-trained language model, this approach further improves performance beyond that of the original text encoder in the teacher model. Note that we presented the other language results in Appendix D.

7 Conclusion

We present a comprehensive study of the knowledge distillation technique in multilingual visual-language model settings. Our study presents the results of design choices that facilitate knowledge transfer to a small model. The experimental results demonstrate that we can decrease the model size from 881M to 433M, where the consistency of the multilingual model is decreased by a margin on retrieval datasets, but it performs similarly on the VQA task. We also present an analysis of performance in the student model and found that, although small models perform lower than the teacher model on retrieval and VQA benchmarks, on tasks that require multilingual consistency (e.g., clustering and ranking), student models can outperform teacher models on these tasks.

Acknowledgement

References

Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In

Proceedings of the 23nd annual ACM symposium on User interface software and technology, pages 333–342.

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 *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR.

Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual clip. In *Proceedings of the thirteenth language resources and evaluation conference*, pages 6848–6854.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

Guanhua Chen, Lu Hou, Yun Chen, Wenliang Dai, Lifeng Shang, Xin Jiang, Qun Liu, Jia Pan, and Wenping Wang. 2023. mCLIP: Multilingual CLIP via cross-lingual transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13028–13043, Toronto, Canada. Association for Computational Linguistics.

Yinpeng Chen, Xiyang Dai, Dongdong Chen, Mengchen Liu, Xiaoyi Dong, Lu Yuan, and Zicheng Liu. 2022. Mobile-former: Bridging mobilenet and transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5270–5279.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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.

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. Association for Computational Linguistics.

Zhiyuan Fang, Jianfeng Wang, Lijuan Wang, Lei Zhang, Yezhou Yang, and Zicheng Liu. 2021. Seed: Self-supervised distillation for visual representation. *arXiv preprint arXiv:2101.04731*.

Geoffrey Hinton. 2015. Distilling the knowledge in a neural network.

Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. 2024. Retrieval-enhanced contrastive vision-text models.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Yash Kant, Abhinav Moudgil, Dhruv Batra, Devi Parikh, and Harsh Agrawal. 2021. Contrast and classify: Training robust vqa models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1604–1613.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Potsawee Manakul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. McCrolin: Multi-consistency cross-lingual training for retrieval question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2780–2793, Miami, Florida, USA. Association for Computational Linguistics.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Lalita Lowphansirikul, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. ConGen: Unsupervised control and generalization distillation for sentence representation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6467–6480, Abu Dhabi, United Arab Emirates. 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 *European conference on computer vision*, pages 740–755. Springer.

Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. 2023. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5607–5612.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. 2019. Relational knowledge distillation.

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 *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

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.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2015. Fitnets: Hints for thin deep nets.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan

Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. 2020. Mobilebert: a compact task-agnostic bert for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170.

Weiting Tan, Kevin Heffernan, Holger Schwenk, and Philipp Koehn. 2023. Multilingual representation distillation with contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1477–1490, Dubrovnik, Croatia. Association for Computational Linguistics.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. 2022. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive representation distillation. *arXiv preprint arXiv:1910.10699*.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kuckreja, et al. 2024. All languages matter: Evaluating lmms on culturally diverse 100 languages. *arXiv preprint arXiv:2411.16508*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Yu Weng, Jun Dong, Wenbin He, Xuan Liu, Zheng Liu, Honghao Gao, et al. 2024. Zero-shot cross-lingual knowledge transfer in vqa via multimodal distillation. *IEEE Transactions on Computational Social Systems*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang,

Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, Boyu Diao, and Yongjun Xu. 2024. Clip-kd: An empirical study of clip model distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15952–15962.

Chuanguang Yang, Helong Zhou, Zhulin An, Xue Jiang, Yongjun Xu, and Qian Zhang. 2022. Cross-image relational knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12319–12328.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Ying Zhao and George Karypis. 2001. Criterion functions for document clustering: Experiments and analysis.

A Experiment setting detail

Model settings Table 6 reports our experiment settings in knowledge distillation methods, and Table 7 reports DR parameters. All parameters are evaluated on the retrieval and VQA benchmarks.
Compute settings All models were trained on a single A100 for ∼24-48 hrs, with the training time depending on the training method.
Training Language The following languages are included in the ImageCaptioning 7M dataset (Training dataset): Afrikaans, Albanian, Amharic, Arabic, Azerbaijani, Bengali, Bosnian, Bulgarian, Catalan, Chinese (Simplified and Traditional), Croatian, Czech, Danish, Dutch, English, Estonian, French, German, Greek, Hindi, Hungarian, Icelandic, Indonesian, Italian, Japanese, Macedonian, Malayalam, Marathi, Polish, Portuguese, Romanian, Russian, Serbian, Slovenian, Spanish, Swahili, Swedish, Tagalog, Telugu, Turkish, Turkmen, Ukrainian, Urdu, Uyghur, Uzbek, and Vietnamese.

Methods	lr	epochs	batch size	warm up steps
FD	$1e^{-5}$	3	64	1000
ED	$1e^{-5}$	3	64	1000
SD	$1e^{-5}$	3	64	1000
MCL	$1e^{-5}$	2	64	1000
DR	$1e^{-4}$	10	64	1000
DR+FD	$1e^{-4}$	10	64	1000

Table 6: Experiment settings in each KD method.

Teacher Model	Student Model	Loss	τ^T	τ^S	K	lr
CLIP-ViT-L/14	DistillBert	DR	0.05	0.07	65536	$1e^{-4}$
CLIP-ViT-L/14	MiniLM	DR	0.05	0.07	65536	$1e^{-4}$
CLIP-ViT-L/14	XLM-R$_{Base}$	DR	0.05	0.07	65536	$1e^{-4}$
SigLIP2-L/16	DistillBert	DR	0.05	0.07	65536	$3e^{-4}$
SigLIP2-L/16	MiniLM	DR	0.05	0.07	65536	$3e^{-4}$
SigLIP2-L/16	XLM-R$_{Base}$	DR	0.05	0.07	65536	$1e^{-4}$

Table 7: Distillation Hyperparameters for the DR method.

B Recall@k results

Table 8 reports $R@5$ and $R@10$ of retrieval benchmarks in various knowledge distillation techniques with SigLIP2-L/16 as a teacher model and XLM-R$_{Base}$ as a student model. The experimental results demonstrate the consistency between $R@1$ (Table 2), $R@5$, and $R@10$, where student models perform similarly to the teacher model on the average score.

C Embedding distribution

Fig 5 reports the embedding distribution of other knowledge distillation methods. As expected, the results of KD models are consistent in that they can perform clustering better than the teacher model, although the retrieval performance is lower than that of the teacher model.

D MRR@k results

Fig 6 reports specific $MRR@K$ in each language of the Multi30k dataset. The MRR is an alternative and confirmation result for the ranking performance (Appendix B). We found that for non-English results, our KD model can perform similar to the teacher model, although the size of text encoder is reduced by half.

E Language performance results

We demonstrate the full retrieval result of each language on benchmarks in Figures 7, 8, 9, 10, 11, and 12.

| | Retrieval (R@5) | | | | | | | |
Methods	Multi30k I2T	Multi30k T2I	COCO I2T	COCO T2I	WIT I2T	xFlickr I2T	XM3600 I2T	AVG.
T: SigLIP2-L/16	90.40	89.80	71.68	42.63	60.83	76.79	72.86	72.14
S: XLM-R$_{Base}$								
+FD	90.37	91.80	58.04	47.82	41.96	83.64	70.61	69.18
+ED	89.73	93.13	57.76	53.08	52.04	85.28	74.03	72.15
+SD	85.77	89.83	51.30	46.41	41.55	80.99	68.08	66.28
+MCL	83.47	86.30	50.24	46.34	39.70	78.72	66.62	
+DR	91.73	93.07	56.28	49.43	54.03	85.84	73.64	72.00
+DR+FD	91.83	92.83	57.42	49.60	54.61	85.73	74.67	72.38
+DR+ED	92.27	93.30	57.48	49.68	54.39	86.00	74.66	72.54
+DR+ED+FD	91.90	93.10	54.64	48.80	53.41	85.33	74.83	71.72
	Retrieval (R@10)							
T: SigLIP2-L/16	94.63	94.10	80.10	54.01	68.21	83.97	78.28	79.04
S: XLM-R$_{Base}$								
+FD	94.37	95.83	67.80	57.78	52.62	89.31	77.51	76.46
+ED	93.83	96.27	69.54	62.70	62.25	90.60	80.15	79.33
+SD	92.13	94.47	61.62	56.82	52.60	87.27	75.55	74.35
+MCL	90.60	92.03	61.68	56.35	49.81	86.05	74.79	
+DR	95.23	95.80	67.02	59.46	62.50	90.65	80.48	78.73
+DR+FD	95.03	95.87	67.34	59.20	63.08	90.86	80.55	78.85
+DR+ED	95.27	96.10	68.22	59.72	62.94	90.81	80.52	79.08
+DR+ED+FD	94.93	95.80	66.70	58.61	62.58	90.40	80.67	78.53

Table 8: Retrieval result (R@5 and R@10) scores on Multi30K, COCO, WIT, xFlickr, and XM3600. The results are from the XLM-RoBERTa base student model trained with knowledge distillation from the SigLIP2-L/16 teacher model.

(a) **XLM-R$_{Base}$-FD**
I2T:70.90±3.97 **T2I:**73.93±2.49 **Purity:**0.335

(b) **XLM-R$_{Base}$-MCL**
I2T:55.50±3.78 **T2I:**64.33±3.13 **Purity:**0.366

(c) **XLM-R$_{Base}$-DRFD**
I2T:71.83±3.22 **T2I:**77.35±2.65 **Purity:**0.359
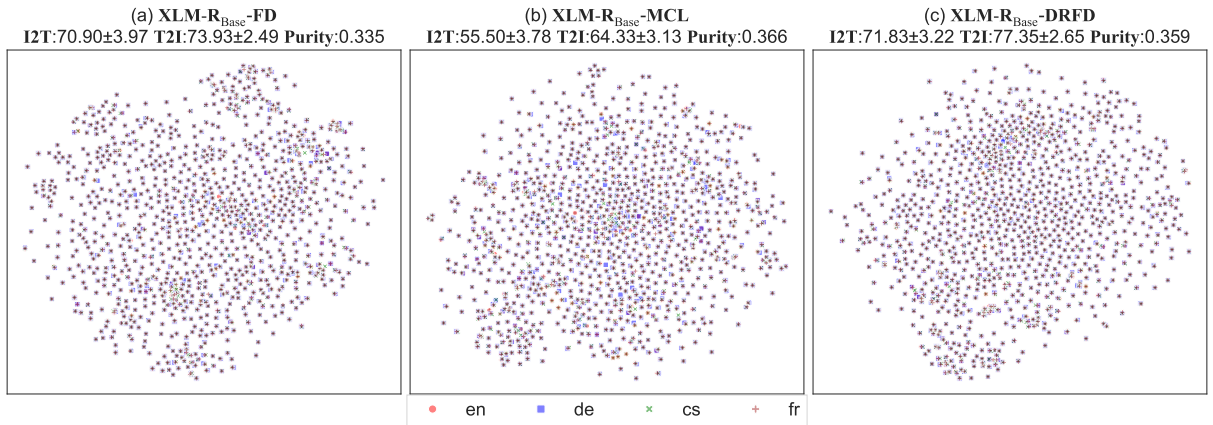
en   de   cs   fr

Figure 5: Embedding distribution of (a) Feature distillation (FD), (b) Multilingual Contrastive Learning (MCL), and in the Multi30k dataset.
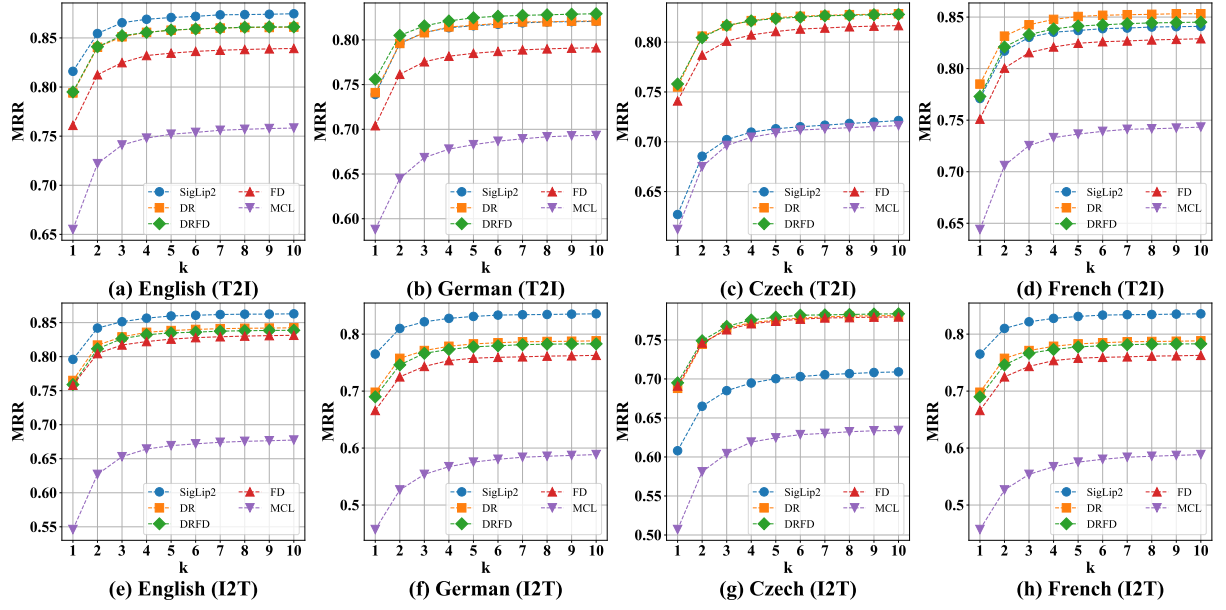
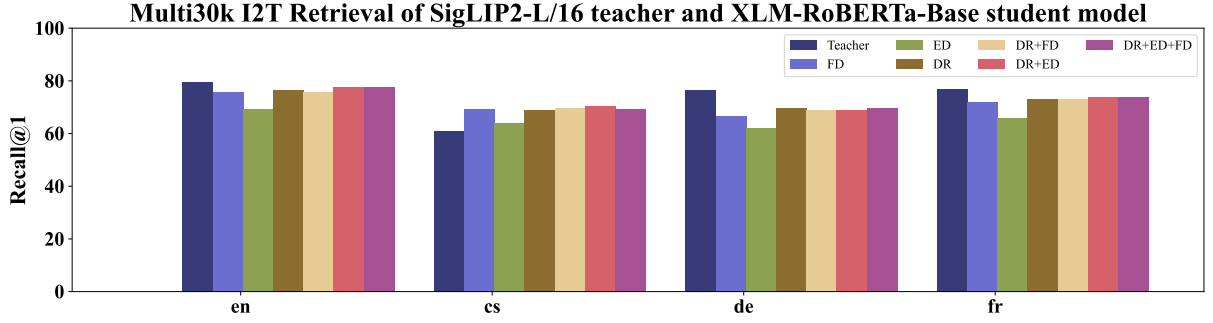Figure 6: Result of MRR@K from Multi30k dataset.

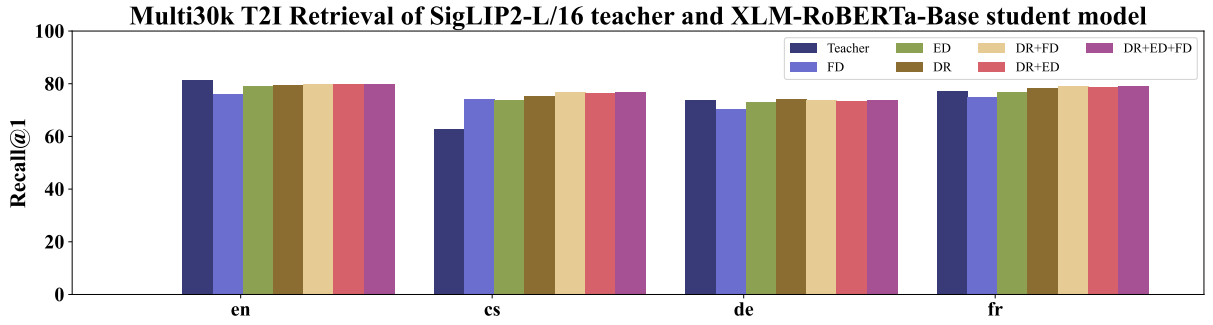Figure 7: Result of recall@1 I2T on the Multi30k dataset.

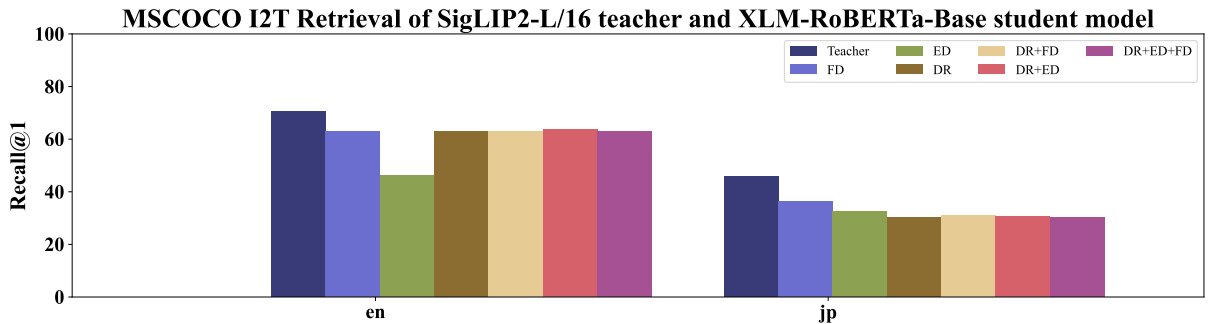Figure 8: Result of recall@1 T2I on the Multi30k dataset.

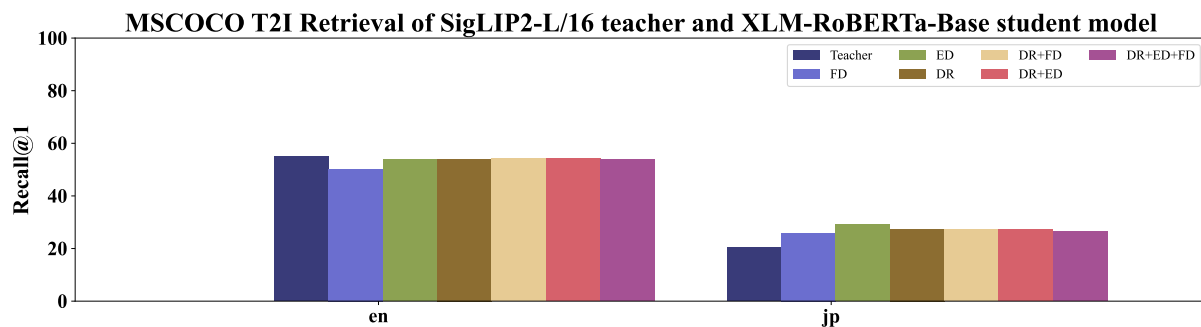Figure 9: Result of recall@1 I2T on the MSCOCO dataset.

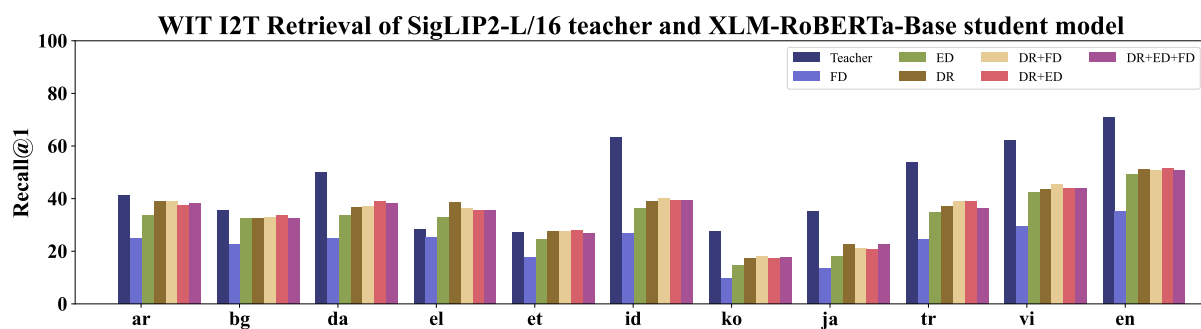Figure 10: Result of recall@1 T2I on the MSCOCO dataset.
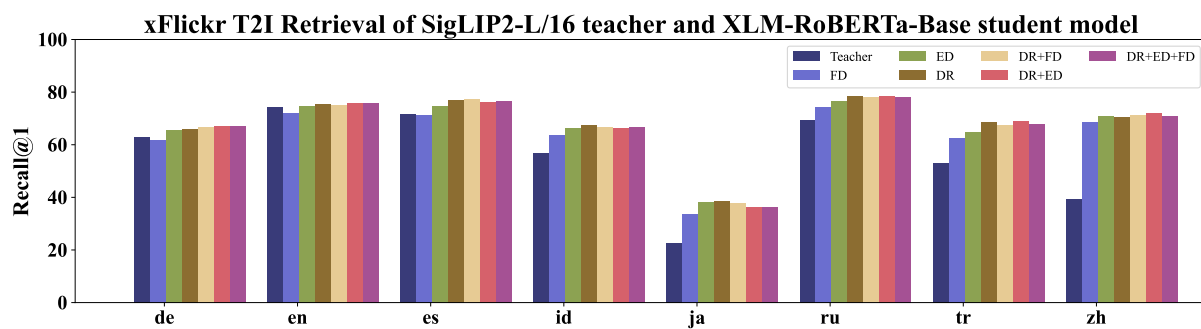

Figure 11: Result of recall@1 on the WIT dataset.

Figure 12: Result of recall@1 on the xFlickr dataset.