MULTILINGUAL-TO-MULTIMODAL (M2M): Unlocking New Languages with Monolingual Text

Anonymous ACL submission

Abstract

Multimodal models have achieved state-of-the-art performance for English language due to abundant high-quality multimodal data (image-text and audio-text). However, the performance for other languages is lower due to limited high-quality multilingual-multimodal data. Current state-of-the-art methods use automatic translations to create and evaluate Multilingual Multimodal models. Meanwhile, the availability of multilingual text data and robust self-supervised methods has grown significantly, leading to powerful multilingual text models. In this work, we leverage the strong multilingual semantic alignment of text models and align them with multimodal models. We demonstrate that learning just a few linear layers can transform multilingual text representations into multimodal text representations that are compatible with the rest of the multimodal model. Our method, M2M, uses only English text data for learning the transformation/alignment. It achieves 95.3% Recall@10 on English language (0.3% higher than the baseline model) and 89.2% Recall@10 averaged across 11 languages (10 of which are unseen during alignment) for the Text-to-Image retrieval task on the XTD dataset. M2M generalizes across architectures, datasets, modalities, and tasks (Image-Text, Audio-Text retrieval, and Cross-lingual Text-to-Image generation). Code, checkpoints, and data will be publicly released[1].

1 Introduction

Humans can naturally align multiple modalities, connecting visual objects with words and sounds. We can associate previously seen objects with their corresponding words in a newly introduced language without explicit supervision or direct mapping between the object and the new word. Instead, this is achieved by implicitly aligning the object, its word in a known language, and the word in the newly introduced language. Existing works (Carlsson et al., 2022; Yan et al., 2024; Koukounas et al., 2024b) explicitly rely on multimodal data in new languages to adapt models like CLIP (Radford et al., 2021) and CLAP (Elizalde et al., 2023) that are primarily trained in English. These multimodal models require large amounts of data for each language, which is often impractical to obtain. In contrast, multilingual text encoders have achieved state-of-the-art (SOTA) performance through the use of abundant text data and self-supervised learning techniques (Devlin et al., 2019; Radford, 2018).

We present a simple method that aligns multilingual and multimodal latent spaces using textual representations as a bridge. Similar to how humans learn, our method doesn't require explicit multimodal signals for each language– English textual data alone is sufficient for alignment. Using robust multilingual text encoders with a simple MSE-like loss function, we achieve performance comparable to models trained on multilingual-multimodal data. We achieve this alignment by learning a projection map (a few linear layers) while keeping the rest of the pretrained model frozen. Maiorca et al. (2024b); Rosenfeld et al. (2022) show the effectiveness of linear projection maps for aligning latent spaces using multimodal English data in classification tasks. Our work extends this approach for multilingual and multimodal latent spaces on retrieval and generative tasks. To summarize our contributions:

1. We propose M2M, a simple alignment method that maps multilingual latent space to multimodal latent space using only monolingual (English) text data. Our empirical results show that M2M is effective across different architectures, evaluation datasets, modalities (image, audio), and tasks (Image-Text & Audio-Text retrieval and Text-to-Image gener-

[1] https://github.com/m2m-acl25/M2M

ation). The method is parameter-efficient, requiring only a few linear layers, and achieves strong alignment even with limited data (∼1K sentences).

2. We create synthetic parallel evaluation datasets for Audio-Text retrieval in 33 languages using test sets from AudioCaps (Kim et al., 2019) (160K samples) and Clotho (Drossos et al., 2019) (172K samples). We also generate 30K MSCOCO captions in 9 new languages (270K samples) for Text-to-Image generation using state-of-the-art translation models.

2 Related Work

Multilingual Multimodal Models. Strong multimodal models like CLIP (Radford et al., 2021) and CLAP (Elizalde et al., 2023; Wu et al., 2022) are typically trained on large amounts of English multimodal data (paired image-text and audio-text data). Extending these models to other languages typically requires explicit training on multilingual-multimodal data—either by training from scratch (Jain et al., 2021) or by finetuning pretrained models (Koukounas et al., 2024b; Yan et al., 2024; Chen et al., 2023; Ye et al., 2024; Li et al., 2023). Some approaches (Carlsson et al., 2022; Chen et al., 2022; Zhai et al., 2021) fine-tune only the text encoders while keeping the image encoder frozen, while Aggarwal and Kale (2020) train projection layers on top of frozen encoders using multimodal English data. Our method uses simple training losses, linear layers, and only English text data. We demonstrate our method's effectiveness for a broad range of multimodal tasks, including Image-Text & Audio-Text retrieval and Text-to-Image generation.

Latent Space Translation is a technique that maps representations between different latent spaces to enable information sharing. Recent research has focused on two main approaches- Using relative representations for latent space alignment (Moschella et al., 2022; Norelli et al., 2022); Creating direct transformation maps (Gower, 1975) between source and target spaces (Maiorca et al., 2024b; Lähner and Moeller, 2024). These approaches have been successfully applied to tasks like cross-modal classification and generative modeling. A further development is the Inverse Relative Projection method (Maiorca et al., 2024a), which converts source representations to relative form before mapping them to a target space, effectively translating monolingual text representations into multilingual text representations. Our approach builds on this foundation by creating a linear mapping between multilingual and multimodal latent spaces. By using English text as a bridge between these spaces, we can create multilingual multimodal models without requiring specialized training data.

3 Methodology

Our method M2M is a simple alignment method that learns a few linear layers to align multilingual latent space with multimodal latent space using English text representations. While we focus on dual-modality multimodal models in this work, our method can extend to models with more than two modalities. Consider a monolingual multimodal model $\mathcal{M}_e$ that supports language e. $\mathcal{M}_e$ consists of individual encoders for each modality. Let $\mathcal{M}_e = (T_e, X_e)$ where T_e is the language e text encoder and X_e represents any other modality encoder (*e.g.* Image, Audio, etc.). We assume representations from both T_e and X_e are already aligned in a shared latent space using paired multimodal data from language e (*e.g.* CLIP, CLAP). Let T_m be a multilingual text encoder. Our goal is to achieve semantic alignment between global representations (sentence-level representations) in the latent space. To achieve this, we learn a projection map $f_{m\to e}$ that transforms multilingual representations from T_m into multimodal representations from T_e using a loss function and language e text-only data. The data must have semantic correspondence with the task/multimodal latent space (*e.g.* image captions for Image-Text retrieval, and audio captions for Audio-Text retrieval). In our method, the projection map—consisting of a few linear layers—is the only learned component, while all encoders (T_e, T_m, X_e) remain frozen. After learning the projection map, we simply replace T_e with $T_{m\to e} = (T_m, f_{m\to e})$ in $\mathcal{M}_e$, resulting in a multilingual multimodal model $\mathcal{M}_m = (T_{m\to e}, X_e)$.

For a sentence s in language e, we extract multimodal text representation $s_e = T_e(s)$ and multilingual text representation $s_m = T_m(s)$. Since both text encoders represent the same sentence s, in an *all-aligned* world, s_e and s_m would be identical. However, s_m and s_e typically differ because they come from different encoders trained with distinct

objectives and datasets. To align s_m and s_e, we learn a projection map $f_{m\to e}$ from s_m to s_e. Here, s_e acts as an anchor to guide latent space translation. We use MSE as our primary loss function.

$$s_{m\to e} = f_{m\to e}(s_m) \quad (1)$$

$$\mathcal{L}_{\text{align}} = \text{MSE}\left(\|s_e\|_2, \|s_{m\to e}\|_2\right) \quad (2)$$

We derive additional supervision from the structure of the target latent space. For a training batch B, let S_e and $S_{m\to e}$ denote the batched sets of multimodal text representations and the corresponding multilingual aligned representations, respectively. We calculate pairwise cosine similarities within the batch for both the target and predicted spaces:

$$C_e = \text{cos\_sim}(S_e, S_e) \quad (3)$$

$$C_{m\to e} = \text{cos\_sim}(S_{m\to e}, S_{m\to e}) \quad (4)$$

where $C_e, C_{m\to e} \in R^{|B|\times|B|}$ are the similarity matrices that capture the structure of the target and predicted spaces, respectively. We minimize the MSE between these similarity matrices:

$$\mathcal{L}_{\text{str}} = \text{MSE}(C_e, C_{m\to e}) \quad (5)$$

This effectively enforces the predicted representations to preserve the structural relationships of the target space. Final loss is a linear combination of both $\mathcal{L}_{align}$ and $\mathcal{L}_{str}$:

$$\mathcal{L} = \lambda * \mathcal{L}_{\text{align}} + \beta * \mathcal{L}_{\text{str}}. \quad (6)$$

We also experiment with other losses such as L1 loss and similarity loss $(1 - \text{cosine}(s_e, s_{m\to e}))$, though these are not as effective as $\mathcal{L}$.

MSE loss helps replace s_e with $s_{m\to e}$ more effectively than contrastive or similarity loss, which only focus on angles between representations. We avoid token/word-level alignment since it emphasizes language structure over semantics. Moreover, implementing the reverse map $f_{e\to m}$ would disrupt the existing alignment between representations from X_e (other modality encoder) and T_e.

4 Experiments and Results

In this section, we empirically demonstrate the effectiveness of our method. We conduct detailed experiments examining the projection map ($f_{m\to e}$) architecture, training loss, and training data scaling in the Image-Text retrieval setting. We evaluate our approach on three tasks: Image-Text Retrieval, Audio-Text Retrieval, and Cross-lingual Text-to-Image generation.

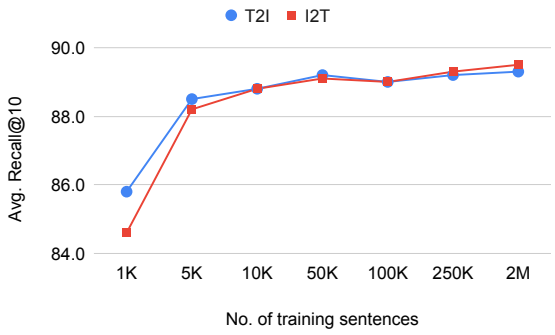


Figure 1: Effect of scaling train data on XTD eval set for M2M-aligned model- Jina-CLIP-v1 × M-MPNET.

4.1 Preliminary Experiments

We investigate the impact of varying number of linear layers (1, 2, 4), adding or removing residual connections (He et al., 2015) in $f_{m\to e}$, and testing different training objectives through ablation studies.

Experimental setup. We primarily use Jina-CLIP-v1 (Koukounas et al., 2024a) as the multimodal model ($\mathcal{M}_e$) and Multilingual MPNET (M-MPNET) (Reimers and Gurevych, 2020) as the multilingual text encoder (T_m). Following (Carlsson et al., 2022), we use a combination of Google Conceptual Captions (GCC) (Sharma et al., 2018), MSCOCO (Lin et al., 2014), and VizWiz (Bigham et al., 2010) as our training dataset to learn $f_{m\to e}$. We remove duplicate sentences and create a N-sentence training split through random sampling. We experiment with various model architectures and training split sizes (Scaling). Unless specified otherwise, we train for 50 epochs using 250K-sentence training size, batch size of 64, AdamW optimizer (Loshchilov, 2017) with a learning rate of 3e-4, weight decay of 1e-2, and a linear learning rate scheduler with 50 warmup steps. All M2M-aligned models are trained on two RTX A5000 24GB Nvidia GPUs. For validation, we use XTD (Aggarwal and Kale, 2020) English image-text pairs, saving the best checkpoint based on the mean of Text-to-Image (T2I) and Image-to-Text (I2T) recall. Both T2I and I2T recalls are averaged across Recall@1,5,10. We evaluate these experiments on Image-to-Text retrieval task using the XTD test dataset.

Result and Analysis. As shown in Table 2, M2M maintains strong performance regardless of the number of linear layers or the presence of residual connections. The performance of our proposed

Models	XTD-T2I												XTD-I2T		XM3600		Multi30K	
	Avg.	de	en	es	fr	it	jp	ko	pl	ru	tr	zh	Avg.	en	T2I	I2T	T2I	I2T
English-only Vision-Language Models																		
E1: CLIP (ViT-L 336px)	35.7	55.4	92.5	64.1	67.0	53.7	18.7	2.7	15.6	5.0	13.2	4.9	43.2	94.1	14.0	23.7	54.9	63.7
E2: Jina-CLIP-v1	37.4	61.5	**95.0**	**67.8**	77.4	58.3	9.8	1.9	16.8	4.4	10.9	**7.5**	39.5	**95.8**	20.3	26.5	58.9	59.6
E3: K-ALIGN	**47.6**	**73.3**	94.0	67.1	**80.0**	**72.8**	**26.2**	**12.6**	**37.6**	**34.0**	**19.1**	7.0	**53.1**	93.8	**22.9**	**31.0**	**67.5**	**70.1**
Multilingual Vision-Language Models Trained on Supervised Multimodal and/or Multilingual Data																		
T1: mUSEM3L	74.9	73.5	85.3	76.7	78.9	78.9	67.8	70.7	71.7	73.6	70.9	76.1	–	–	–	–	–	–
T2: MCLIP-ST	76.4	78.7	88.5	78.2	79.8	79.3	68.6	63.1	75.6	74.7	74.4	79.4	78.6	90.4	48.7	60.6	80.7	83.4
T3: ALIGN-Base	82.2*	–	–	88.8	–	87.9	–	76.6	79.8	82.3	73.5	86.5	–	–	–	–	–	–
T4: MURAL-Large	90.2*	–	–	92.9	–	91.8	–	88.1	91.0	87.2	89.5	89.7	–	–	–	–	–	–
T5: LABSE ViT-L/14	87.2	89.6	91.6	89.5	89.9	90.1	73.9	80.8	89.8	85.5	89.8	88.9	90.8	94.9	73.2	83.6	90.9	93.7
T6: XLM-R-L ViT-B/32	88.0	88.7	91.8	89.1	89.4	89.8	81.0	82.1	91.4	86.1	88.8	89.3	89.9	91.7	75.2	84.5	89.2	91.0
T7: XLM-R ViT-L/14	89.0	90.6	92.4	91.0	90.0	91.1	81.9	85.2	91.3	85.8	90.3	89.7	92.2	94.5	76.4	85.0	92.2	**94.4**
T8: XLM-R-L ViT-B/16+	92.0	**93.0**	95.0	93.6	93.1	93.1	84.2	89.0	94.4	90.0	93.0	94.0	**93.2**	**96.1**	**81.8**	**87.1**	**93.9**	94.2
T9: Jina-CLIP-v2	**92.6**	92.5	92.8	88.9	**95.5**	93.2	**94.1**	90.6	**94.9**	90.7	**93.5**	91.4	**93.2**	92.7	81.1	85.7	93.8	94.0
T10: AltCLIP$_{M9}$	**93.7***	–	**95.4**	**94.1**	92.9	**94.2**	91.7	**94.4**	–	**91.8**	–	**95.1**	–	–	–	–	–	–
M2M-aligned Multilingual Multimodal models using English-only Text data																		
M1: Jina-CLIP-v1 × LaBSE	82.4	82.5	86.4	83.7	84.4	84.5	76.2	80	84.5	80.0	80.7	83.0	80.1	87.3	62.8	65.6	78.7	75.2
M2: Jina-CLIP-v1 × M-MiniLM	86.5	87.5	94.1	88.2	88.0	87.4	80.6	74.8	89.2	85.0	86.2	**90.1**	84.9	93.8	57.7	64.3	87.8	85.7
M3: Jina-CLIP-v1 × JinaTextV3	87.8	**91.0**	**95.3**	89.5	**90.1**	**91.2**	80.4	80.1	90.2	85.6	87.4	84.9	87.5	94.7	**67.0**	72.2	87.9	87.2
M4: Jina-CLIP-v1 × M-MPNET	**89.2**	90.9	94.4	**91.1**	89.5	90.8	**82.4**	**85.4**	**90.6**	**87.1**	**88.9**	**90.1**	**89.3**	**95.6**	66.4	**72.9**	90.0	89.7
M5: CLIP × M-MPNET	84.2	85.4	91.0	85.6	85.1	85.8	77.8	80.6	84.7	81.6	84.5	83.9	85.9	93.7	55.5	66.9	**90.6**	**92.0**
M6: K-ALIGN × M-MPNET	86.8	87.5	93.0	89.7	87.8	88.3	78.7	83.0	88.6	83.2	87.0	87.5	86.1	94.2	59.1	68.5	90.4	90.0

Table 1: Comparison of M2M-aligned model performance with English and Multilingual CLIP-like models using Recall@10 across datasets. Results include reported XTD-T2I numbers for T1, T3-T8, T10 and rest are computed using available checkpoints. * denotes average is computed over only supported languages.

Loss	Linear layers	Skip Conn.	Avg.	de	en	es	fr	it	jp	ko	pl	ru	tr	zh
MSE	2	No	88.7	89.0	95.4	89.9	89.3	89.5	82.0	85.3	90.2	85.7	89.5	90.1
MSE	2	Yes	88.8	88.8	95.0	90.1	89.6	90.2	82.0	85.4	90.5	85.6	89.2	89.9
$\lambda_1 \cdot \mathcal{L}_{align} + \beta_1 \cdot \mathcal{L}_{str}$	2	No	89.3	89.4	95.6	90.7	89.3	90.6	82.3	85.2	**91.5**	86.5	**90.4**	90.4
$\lambda_1 \cdot \mathcal{L}_{align} + \beta_1 \cdot \mathcal{L}_{str}$	2	Yes	89.2	89.2	95.2	91.0	89.4	90.3	82.6	85.8	91.0	86.5	89.6	90.4
$\lambda_1 \cdot \mathcal{L}_{align} + \beta_1 \cdot \mathcal{L}_{str}$	4	No	89.3	**89.5**	**95.7**	**91.1**	89.6	**90.8**	82.5	**86.2**	90.9	86.5	90.0	90.0
$\lambda_1 \cdot \mathcal{L}_{align} + \beta_1 \cdot \mathcal{L}_{str}$	1	No	**89.4**	89.2	95.4	91.0	**89.7**	90.7	**82.9**	85.5	90.8	**86.9**	**90.4**	**90.5**
$\lambda_2 \cdot \mathcal{L}_{align} + \beta_1 \cdot \mathcal{L}_{str}$	2	No	88.8	89.1	94.8	90.6	89.6	90.1	81.9	85.2	90.7	86.1	89.3	89.2
Similarity Loss	2	No	88.6	89.0	95.5	89.6	89.2	89.8	81.7	84.9	90.4	85.6	89.1	90.1
L1	2	No	84.4	85.9	94.7	86.6	85.3	87.2	76.3	78.8	84.8	78.8	84.3	85.7

Table 2: Comparison of Recall@10 across different training losses, linear layers, and residual connections (Skip Conn.) for M2M-aligned Jina-CLIP-v1×M-MPNET on XTD Image-to-Text retrieval. $\lambda_1 = 44, \lambda_2 = 1, \beta_1 = 1$.

training objective (eq. 6) surpasses alternative approaches for Image-to-Text retrieval across 11 languages, with improvements in Avg. Recall@10 of absolute 0.7% over MSE loss, 0.8% over similarity loss, and 5% over L1 loss. Text-to-Image retrieval task shows similar results (see Appendix D). Assigning a higher weight for $\mathcal{L}_{align}$ ($\lambda = 44$) compared to $\mathcal{L}_{str}$ ($\beta = 1$) yields a 0.6% gain in Avg. Recall@10 versus equal weighting ($\lambda = 1, \beta = 1$).

Based on the optimal configuration from Table 2, we conduct data scaling experiments for the projection map $f_{m \to e}$ with 2 linear layers, no residual connection, and parameters $\lambda = 44, \beta = 1$. We train with loss $\mathcal{L}$ (eq. 6) using training splits containing 1K, 5K, 10K, 50K, 100K, 250K, and 2M sentences. Figure 1 demonstrates that M2M achieves 85.8% Avg. Recall@10 (across 11 languages) with just 1,000 English sentences, without any multilingual or multimodal data. We observed diminishing returns from scaling beyond 250K sentences. Increasing the data fourfold to 2M sentences yielded only minimal improvements of 0.1% (T2I) and 0.2% (I2T) in Avg. Recall@10 (see Appendix E for details). Therefore, all experiments use the 250K-sentence training split by default. Although a single layer had higher performance for map $f_{m \to e}$, we opt for 2 linear layers (∼1M parameters) as the optimal number since there can be a mismatch of latent dimensions between multilingual and multimodal spaces. We use the first linear layer to project multilingual representations to match the dimension of multimodal space.

4.2 Image-Text Retrieval

Experimental setup. We experiment with several English multimodal models ($\mathcal{M}_e$): CLIP (Radford et al., 2021), Jina-CLIP-v1 (Koukounas et al., 2024a), and KakaoBrain-ALIGN (K-ALIGN) (Yoon et al., 2022), along with multilingual text encoders (T_m): LaBSE (Feng et al., 2020), Multilingual MPNET (M-MPNET) (Reimers and Gurevych, 2020), Multilingual MiniLM (M-MiniLM) (Reimers and Gurevych, 2020), and Jina-embeddings-v3 (Jina-Text-v3) (Sturua et al., 2024). Models aligned using our method are denoted as `Multimodal-Model`×`Multilingual-Model`. We compare M2M-aligned models against English-only Vision-Language models (CLIP, Jina-CLIP-v1, K-ALIGN) and existing Multilingual Multimodal Models (MMMs): mUSEM3L (Aggarwal and Kale, 2020), Multilingual CLIP from SentenceTransformer Library[2] (MCLIP-ST) (Reimers and Gurevych, 2020), MURAL-Large (Jain et al., 2021), ALIGN-Base (Jia et al., 2021) reported by MURAL, (Carlsson et al., 2022)'s LaBSE ViT-L/14, XLM-R-Large ViT-B/32, XLM-R ViT-L/14, XLM-R-Large ViT-B/16+, Jina-CLIP-v2 (Koukounas et al., 2024b), and AltCLIP$_{M9}$ (Chen et al., 2022). Languages supported by these models are listed in Appendix C.

Evaluation. We evaluate using three multilingual datasets: XTD (11 languages) (Aggarwal and Kale, 2020), which includes MIC (Rajendran et al., 2016) (de, fr) and STAIR Captions (Yoshikawa et al., 2017) (jp); XM3600 (36 languages) (Thapliyal et al., 2022); and Multi30K (4 languages) (Elliott et al., 2016) (Elliott et al., 2017) (Barrault et al., 2018). Following previous works (Aggarwal and Kale, 2020; Jain et al., 2021; Carlsson et al., 2022), we evaluate using Recall@10 with cosine similarity as the ranking score. For XTD, we report Text-to-Image retrieval scores across all languages along with Avg. Recall@10. For XM3600, Multi30K, and Image-to-Text retrieval task, we only report the mean Recall@10 score across all languages present in the dataset with per language score in Appendix F.

Results & Analysis. For the XTD T2I task, M2M-aligned Jina-CLIP-v1 × M-MPNET model (row M4 in Table 1) outperforms several MMMs trained on multimodal and/or multilingual paired data (rows T1-T3, T5-T7). For English, our Jina-CLIP-v1×Jina-Text-v3 model (row M3) outperforms all English-only baselines (rows E1-E3). For subsequent comparisons, we use Jina-CLIP-v2 as SOTA which has the best performance averaged across all languages.

On the XTD dataset, our best M2M-aligned model (row M4) performs 3.4% lower on T2I and 3.9% lower on I2T compared to SOTA. This performance gap is expected, as models like Jina-CLIP-v2 are explicitly trained on massive amounts of multilingual-multimodal i.e. ∼400M non-English image-text pairs from CommonPool (Gadre et al., 2023) and 1.2M multilingual synthetic captions. For the Multi30K dataset, we observe a simlar performance gap of 3.3% for T2I and 2.4% for I2T. However, for XM3600, this gap widens to 14.2% for I2T and 14.8% for T2I. We speculate this is due to it's larger retrieval space (Multi30K and XTD have 1K instances in test set, compared to XM3600 that contains 3,600 images and ∼7K captions). When considering only model-supported languages, this gap narrows to 10.1% for I2T and 13.2% for T2I. Detailed performance metrics for XM3600 and Multi30K's supported languages are available in the Appendix F.

4.3 Audio-Text Retrieval

Experimental Setup. We use LAION-CLAP (Wu et al., 2022) as the Audio-Text multimodal model ($\mathcal{M}e$) and align it with the M-MPNET (T_e). We experiment with two variants: 1) CLAP-HTSAT-fused (trained on AudioCaps (Kim et al., 2019), Clotho (Drossos et al., 2019), and LAION-Audio-630k dataset (Wu et al., 2022)) and 2) CLAP-General (trained on additional speech and music data). For alignment, we use English text of audio-caption datasets: AudioCaps, Clotho, and WavCaps (Mei et al., 2023). We use the AudioCaps validation set to save the best checkpoint.

Synthetic Evaluation Datasets. Due to the lack of multilingual audio-text evaluation datasets, we extend AudioCaps (4875 captions) and Clotho (5225 captions) test sets to 33 new languages using machine translation models. We use English-to-Indic translation model from IndicTrans2 (Gala et al., 2023) for 11 Indic languages [3] and Aya-23-35B (Aryabumi et al., 2024) for 22 other languages [4]. Based on the results reported by Aya-23-35B on the FLoRes-200 test set (Costa-jussà et al., 2022) and manual spot-check, we assume that the

[2] https://www.sbert.net/

[3] bn, gu, hi, kn, ml, mr, ne, pa, ta, te, ur

[4] ar, zh-Hans, zh-Hant, cs, nl, fr, de, el, he, id, it, ja, ko, fa, pl, pt, ro, ru, es, tr, uk, vi

Models	AudioCaps				Clotho			
	T2A		A2T		T2A		A2T	
	Avg.	en	Avg.	en	Avg.	en	Avg.	en
English-only LAION-CLAP Models								
CF: HTSAT-Fused	–	70.3/82.5*	–	74.4/88.0*	–	**49.9/55.4***	–	**60.9/66.9***
CG: General	–	**83.4**	–	**89.7**	–	49.3	–	58.1
M2M-aligned Multilingual CLAP models								
CM1: CF × M-MPNET	42.8/47.1†	62.6	51.3/55.3†	63.5	33.2/36.4†	46.7	**39.7/42.8†**	49.9
CM2: CG × M-MPNET	**48.3/54.2†**	**77.2**	**60.8/65.9†**	**81.5**	**33.3/36.7†**	**47.8**	39.6/**42.8†**	**50.6**

Table 3: Performance comparison of Audio-Text Models on AudioCaps and Clotho datasets using Recall@10 for Text-to-Audio (T2A) and Audio-to-Text (A2T) retrieval, averaged across 34 languages. * denotes reported numbers from Wu et al. (2022) and rest are computed from checkpoints. † represents Avg. over supported languages.

translations for the 22 languages are of reasonably high quality. Additionally, we use the FLoRes-200 test set to find the optimal prompt to be used for obtaining the translations. To assess the translation quality for Indic languages, we back-translate the translations to English using the IndicTrans2 (indic-to-en). Across 11 Indic languages, we observe a mean spBLEU (Post, 2018) score of 48.7 and chrF++ (Popović, 2017) score of 63.6 for the AudioCaps test set. For Clotho test set, the mean spBLEU is 47.4 and mean chrF++ is 59.6. Additional details about dataset license and translation quality assessment are discussed in the Appendix B, G. Due to lack of comparable multilingual baselines and test sets, we report Recall@10 metric for our method only on our synthetic multilingual test sets. Language-wise Recall@10 are reported in Appendix H for both AudioCaps and Clotho.

Results & Analysis. Table 3 demonstrates our method's effectiveness in generalizing across modalities beyond images. For English, our method performs below the state-of-the-art by 6.2% on Text-to-Audio retrieval (T2A) and 8.2% on Audio-to-Text retrieval (A2T) using the AudioCaps test set, and by 2.1% (T2A) and 7.5% (A2T) on the Clotho test set.

To investigate this drop, we compute Text-to-Text (T2T) Recall@10 on XM3600 (image-text) and AudioCaps (audio-text) test sets as these datasets contain multiple captions for each image/audio. M-MPNET (multilingual text encoder) achieves a T2T Recall@10 of 62.1%, comparable to Jina-CLIP-v1 (image-text model) at 63.8% on the image-text test set. However, the same M-MPNET achieves 73.8%, i.e. significantly lower than CLAP-general's (audio-text model) at 80.2%. We speculate that M-MPNET excels in image-caption encoding but underperforms in audio-caption encoding in general.

Additionally, our qualitative analysis reveals strong semantic alignment between audio and text representations. For example, when given the query "A man speaks with some clicks and then loud long scrapes", the top three retrieved audio captions were: 1) "Sanding and filing then a man speaks", 2) "A man speaks with some clicking and some sanding", and 3) "A man speaks with a high-frequency hum with some banging and clanking". Although the ground truth audio appeared at rank 10, its captions closely matched those of the top retrieved results: "A man talking as metal clacks followed by metal scraping against a metal surface", and "A man is speaking followed by saw blade noises". This semantic overlap between ground truth and retrieved audio-captions indicates strong retrieval performance. Please refer to Appendix H for more details and additional examples.

4.4 Cross-lingual Text-to-Image Generation

Our method is also agnostic of tasks and extends to generative tasks like Text-to-Image generation. Since M2M aligns sentence-level representations (CLS) of models, we experiment with a Text-to-Image generation model that utilizes CLS of the text, namely FLUX.1-dev (FLUX) (Labs, 2024). FLUX is a 12B parameter model chosen due to its public availability, competitive performance (Yang et al., 2024), and use of CLIP text encoder (CLS conditioning). Apart from CLIP, FLUX contains a T5 encoder (Raffel et al., 2020) for token conditioning. To learn the projection map $f_{m \to e}$, we select the M-MPNET model [5] (T_m) and the CLIP encoder from FLUX (T_e). Since FLUX uses two text encoders (CLIP and T5), we experiment with four settings:

[5]We qualitatively analyzed 100 generated images for FLUX × LaBSE and FLUX × M-MPNET aligned models and found the latter generated better quality images.

(a) FLUX (en) (b) FLUX CLIP (en) (c) Ours (el) (d) Ours (fa) (e) Ours (fr) (f) Ours (he)

(g) FLUX-T5 (en) (h) Ours (en) (i) Ours (hi) (j) Ours (id) (k) Ours (ko) (l) Ours (es)

Figure 2: Images generated by FLUX text-to-image model using the prompt "The city bus is traveling down the road" in multiple languages. Our M2M-aligned model produces similar quality images compared to baseline FLUX (both T5 and CLIP encoders), FLUX-T5 and FLUX-CLIP models.

1. FLUX: Generate images using the same input text for both CLIP and T5 encoders

2. FLUX-CLIP: Use input text for CLIP encoder and a generic text prompt for T5 encoder: "A photo of: " [6]

3. FLUX-T5: Use input text for T5 encoder and a generic text prompt for CLIP encoder: "A photo of: "

4. FLUX × M-MPNET: Use input text for M2M-aligned M-MPNET encoder and a generic text prompt for T5 encoder: "A photo of: "

FLUX × M-MPNET represents our proposed M2M-aligned zero-shot Cross-lingual Text-to-Image generation model.

Training Setup & Evaluation. We follow the training settings and dataset described in Section 4.1, but train for 10 epochs without a validation set, using bfloat16 precision and MSE loss (instead of loss in eq. 6) over unnormalized representations. Using $\mathcal{L}_{str}$ with MSE leads to significant degradation. The identical mapping of representations is more important for generation task than the structural similarity between latent spaces. We generate 512×512 resolution images using 3.5 guidance scale, 10 inference steps, and a fixed seed. Following evaluation protocols from previous works (Ramesh et al., 2021; Rombach et al., 2021; Saharia et al., 2022), we randomly sample 30K captions from the MSCOCO2014 (Lin et al., 2014) validation set. For multilingual evaluation, we follow the process outlined in Synthetic Evaluation datasets–section 4.3 and generate parallel captions in 9 new languages [7]. We evaluate performance using FID (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016).

Results and Analysis. Table 4 shows FLUX × M-MPNET achieves high Inception score of 31.81 (averaged over all languages) including English (35.9±0.57), surpassing trained models such as Latent Diffusion Model (LDM) (Rombach et al., 2021) (30.29±0.42), CogView (Ding et al., 2021) (18.2), and LAFITE (Zhou et al., 2022) (26.02). Both, FLUX-CLIP and FLUX × M-MPNET show a poor FID score of 40.9 and 43.4 (averaged over all languages) respectively, while FLUX and FLUX-T5 have a significantly better and same FID score of 23.4. The same lower FID between FLUX and FLUX-T5 indicates that the FLUX model relies heavily on T5-token representations, and can generate high-quality images without any signals from the CLIP encoder. Since our method is not using the FLUX model as intended (with both CLIP and T5 encoder), FLUX × M-MPNET generated images have suboptimal quality and are less faithful to the conditioned multilingual text (*e.g.* missing objects due to lost signal from T5 encoder).

[6] We qualitatively analyzed 100 generated images FLUX-CLIP, FLUX × M-MPNET setting with prompts- "*An image of:* ","*A picture of:* ", "*A photo of:* " and found the prompt- "*A photo of:* " generated better quality images.

[7] fr, el, he, id, ko, fa, ru, es, hi. We use IndicTrans2 for hi, and AYA-24-35B for remaining languages.

(a) Ours (hi) (b) Ours (fr) (c) Ours (hi) (d) Ours (fr) (e) Ours (hi) (f) Ours (fr)

(g) Ours (ru) (h) Ours (fa) (i) Ours (ru) (j) Ours (fa) (k) Ours (ru) (l) Ours (fa)

"A green-themed photo of: " "A cartoon photo of: " "A Van Gogh-style photo of: "

Figure 3: Images generated from multilingual translations of input prompt: "The city bus is traveling down the road" using FLUX × M-MPNET model, with theme prompts in T5 encoder to enhance image quality and style.

Models	Inception Score (↑)									
	en	fr	el	he	id	ko	fa	ru	es	hi
FLUX	42.3±0.81	-	-	-	-	-	-	-	-	-
FLUX-T5	42.1±0.64	-	-	-	-	-	-	-	-	-
FLUX-CLIP	33.4±0.57	-	-	-	-	-	-	-	-	-
FLUX × M-MPNET	35.9±0.57	32.7±0.80	29.9±0.66	29.9±0.45	34.3±0.76	30.2±0.51	32.5±0.74	28.6±0.63	32.8±0.50	31.3±0.46

Table 4: Inception score for MSCOCO-30K on 512×512 images (10 inference steps; guidance scale = 3.5).

Despite this limitation, our qualitative analysis reveals diverse high-quality, slightly low-fidelity cross-lingual image generations shown in Figure 2. For FLUX-CLIP and FLUX × M-MPNET, we also notice hallucinated images–generated images that are unrelated to the given text but remain coherent and well-formed, potentially misaligned. These images are neither random noise nor contain misplaced/distorted object features. We suspect signal loss from the T5 encoder due to generic prompt input may lead to these hallucinations, potentially resulting in higher FID scores compared to FLUX and FLUX-T5. To alleviate these issues, we can simply add missing objects, style, theme, etc. in the prompt to T5 encoder, as shown in Figure 3. Please refer to Appendix I for more details, examples, and language-wise FID scores.

5 Conclusion

In this work, we introduced M2M —an efficient alignment method that transforms multilingual latent space into multimodal latent space using only a few linear layers and English text data. Unlike existing methods that require extensive multilingual and multimodal datasets, our approach significantly reduces resource requirements while maintaining robust performance across diverse tasks and modalities. Our method demonstrates consistent generalization across training strategies, datasets, modalities, and tasks, achieving a 95.3% Recall@10 for English and a strong zero-shot multilingual performance with an average Recall@10 of 89.2% across 11 languages on XTD-T2I retrieval. Through both qualitative and quantitative analysis, we show our method's effectiveness for Image-Text & Audio-Text retrieval, and Text-to-Image generation. To facilitate future research, we release our synthetic evaluation datasets: AudioCaps & Clotho in 33 new languages and MSCOCO 30K captions in 9 new languages, providing a unified framework for benchmarking multilingual performance on multimodal tasks. While these results are promising, there remains room for improvement, particularly in exploring token-level alignment. We hope our work encourages approaches that leverage implicit alignment between languages and modalities, rather than relying solely on additional data to enhance performance on multimodal tasks.

Limitations

Need for local alignment. While our method performs well compared to trained multilingual multimodal models in global-representation (sentence) space, we need to develop alignment at the local-representation (token) level. Tasks like Text-to-Image Generation and cross-lingual skill transfer would benefit significantly from fine-grained signals alongside high-level semantics. Our current method does not support local alignment, and we present this as an opportunity for future research.

Joint Cross-modal Representations. Our work effectively aligns multilingual and multimodal representations from dual encoder models, where each modality is encoded individually. Joint cross-modal encoders generate representations by combining multiple modality representations through shared architectural components. The effectiveness of our method for joint cross-modal representations remains to be explored.

Lack of Human-verified multilingual-multimodal evaluation set. Finding high-quality standard multilingual evaluation sets for Audio-Text retrieval and Text-to-Image Generation tasks is challenging. To address this, we curated synthetic parallel evaluation data for AudioCaps (160K samples), Clotho (172K samples), and MSCOCO-30K (270K samples). Due to the large scale of the data, human verification of the translated captions was not feasible for us. While we use objective metrics like spBLEU and chrF++ to ensure dataset quality, these measures alone are not sufficient, and without human verification, some errors may persist in the evaluation dataset.

References

Pranav Aggarwal and Ajinkya Kale. 2020. Towards zero-shot cross-lingual image retrieval. *arXiv preprint arXiv:2012.05107*.

Nahid Alam, Karthik Reddy Kanjula, Surya Guthikonda, Timothy Chung, Bala Krishna S Vegesna, Abhipsha Das, Anthony Susevski, Ryan Sze-Yin Chan, SM Uddin, Shayekh Bin Islam, et al. 2024. Maya: An instruction finetuned multilingual multimodal model. *arXiv preprint arXiv:2412.07112*.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr F. Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, A. Ustun, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *ArXiv*, abs/2405.15032.

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

Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Rob Miller, Robin Miller, Aubrey Tatarowicz, Brandyn Allen White, Samuel White, and Tom Yeh. 2010. Vizwiz: nearly real-time answers to visual questions. *Proceedings of the 23nd annual ACM symposium on User interface software and technology*.

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, Marseille, France. European Language Resources Association.

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.

Zhongzhi Chen, Guangyi Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Yu Wu. 2022. Altclip: Altering the language encoder in clip for extended language capabilities. *ArXiv*, abs/2211.06679.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou

Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2019. Clotho: an audio captioning dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Matthew Cer, N. Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. In *Annual Meeting of the Association for Computational Linguistics*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander J. Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alexandros G. Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. 2023. Datacomp: In search of the next generation of multimodal datasets. *ArXiv*, abs/2304.14108.

Jay P. Gala, Pranjal A. Chitale, AK Raghavan, Varun Gumma, Sumanth Doddapaneni, M. AswanthKumar, Janki Atul Nawale, Anupama Sujatha, Ratish Puduppully, Vivek Raghavan, Pratyush Kumar, Mitesh M. Khapra, Raj Dabre, and Anoop Kunchukuttan. 2023. Indictrans2: Towards high-quality and accessible machine translation models for all 22 scheduled indian languages. *Trans. Mach. Learn. Res.*, 2023.

John C Gower. 1975. Generalized procrustes analysis. *Psychometrika*, 40:33–51.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

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

Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. Mural: Multimodal, multitask retrieval across languages. *ArXiv*, abs/2109.05125.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Mart'inez, Saahil Ognawala, Susana Guzman, Maximilian Werk, Nan Wang, and Han Xiao. 2024a. Jina clip: Your clip model is also your text retriever. *ArXiv*, abs/2405.20204.

Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Gunther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. 2024b. jina-clip-v2: Multilingual multimodal embeddings for text and images. *ArXiv*, abs/2412.08802.

Black Forest Labs. 2024. Flux. `https://github.com/black-forest-labs/flux`. Accessed: 2025-02-11.

Zorah Lähner and Michael Moeller. 2024. On the direct alignment of latent spaces. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, pages 158–169. PMLR.

Yaoyiran Li, Ching-Yun Chang, Stephen Rawls, Ivan Vulić, and Anna Korhonen. 2023. Translation-enhanced multilingual text-to-image generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9174–9193, Toronto, Canada. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. 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*.

I Loshchilov. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Valentino Maiorca, Luca Moschella, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. 2024a. Latent space translation via inverse relative projection. *arXiv preprint arXiv:2406.15057*.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. 2024b. Latent space translation via semantic alignment. *Advances in Neural Information Processing Systems*, 36.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. 2022. Relative representations enable zero-shot latent space communication. *ArXiv*, abs/2209.15430.

Antonio Norelli, Marco Fumero, Valentino Maiorca, Luca Moschella, Emanuele Rodolà, and Francesco Locatello. 2022. Asif: Coupled data turns unimodal models to multimodal without training. *ArXiv*, abs/2210.01738.

Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. 2020. High-fidelity performance metrics for generative models in pytorch. Version: 0.3.0, DOI: 10.5281/zenodo.4957738.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

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. *Journal of machine learning research*, 21(140):1–67.

Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 171–181, San Diego, California. Association for Computational Linguistics.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092.

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*. Association for Computational Linguistics.

Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Elan Rosenfeld, Preetum Nakkiran, Hadi Pouransari, Oncel Tuzel, and Fartash Faghri. 2022. Ape: Aligning pretrained encoders to quickly learn aligned multimodal representations. *ArXiv*, abs/2210.03927.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *ArXiv*, abs/1606.03498.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Annual Meeting of the Association for Computational Linguistics*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

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.

Yusong Wu, K. Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Zhiyong Yan, Heinrich Dinkel, Yongqing Wang, Jizhong Liu, Junbo Zhang, Yujun Wang, and Bin Wang. 2024. Bridging language gaps in audio-text retrieval. *ArXiv*, abs/2406.07012.

Chenglin Yang, Celong Liu, Xueqing Deng, Dongwon Kim, Xing Mei, Xiaohui Shen, and Liang-Chieh Chen. 2024. 1.58-bit flux. *arXiv preprint arXiv:2412.18653*.

Fulong Ye, Guang Liu, Xinya Wu, and Ledell Wu. 2024. Altdiffusion: A multilingual text-to-image diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 6648–6656.

Boogeo Yoon, Youhan Lee, and Woonhyuk Baek. 2022. Coyo-align. `https://github.com/kakaobrain/coyo-align`.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 417–421, Vancouver, Canada. Association for Computational Linguistics.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2021. Lit: Zero-shot transfer with locked-image text tuning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18102–18112.

Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. 2022. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17907–17917.

A Potential Risks

There has been investigation of various biases (gender, race, etc.) for multimodal models primarily in English language. Our method extends the capability of multimodal models to many languages including low resource languages. However, there have been very few works to detect and mitigate biases for these languages. Additionally, since we use English language as anchor it is possible that the biases present in English multimodal model can manifest in the resulting multilingual multimodal model.

B Model & Data License

All models that are taken from sentence-transformers[8] library (Multilingual CLIP (MCLIP-ST), Multilingual MPNET (M-MPNET), Multilingual MiniLM (M-MiniLM)), LaBSE, KakaoBrain-ALIGN, Jina-CLIP-v1, and LAION-CLAP (CLAP-General, CLAP-HTSAT-Fused) are under Apache License 2.0. For model FLUX.1-dev, generated outputs can be used for personal, scientific, and commercial purposes as described in the FLUX.1 [dev] Non-Commercial License. Multilingual CLIP (Carlsson et al., 2022), OpenAI-CLIP, and IndicTrans2 are under MIT License. Jina-CLIP-v2, Jina-embeddings-v3, AYA-23-35B are under CC-by-NC-4.0. Use of any combination of the models aligned using our method must adhere to the license of all individual models.

We release our extended datasets in new languages for AudioCaps, Clotho, and MSCOCO2014-30K under CC-By-NC-4.0 License, adhering to source dataset licenses and models used to generate data (AudioCaps- MIT License, Clotho- Tampere University License (non-commercial with attribution), MSCOCO-CC-By-4.0).

C List supported languages for multilingual and/or multimodal models

Different multilingual text encoder and multilingual CLIP models support different languages. For fairer comparison, we also report metrics averaged on model-supported languages (e.g. Table 3 and Table 11). Table 5 shows a list of models and their supported languages.

D Preliminary experiments on Text-to-Image Retrieval

Table 6 shows our method outperforms all other training objectives on Text-to-Image retrieval for XTD dataset. The impact of high λ is less significant for Text-to-Image retrieval (equal performance

[8] `https://www.sbert.net/`

Models	Supported languages
LaBSE (Feng et al., 2020)	af, ht, pt, am, hu, ro, ar, hy, ru, as, id, rw, az, ig, si, be, is, sk, bg, it, sl, bn, ja, sm, bo, jv, sn, bs, ka, so, ca, kk, sq, ceb, km, sr, co, kn, st, cs, ko, su, cy, ku, sv, da, ky, sw, de, la, ta, el, lb, te, en, lo, tg, eo, lt, th, es, lv, tk, et, mg, tl, eu, mi, tr, fa, mk, tt, fi, ml, ug, fr, mn, uk, fy, mr, ur, ga, ms, uz, gd, mt, vi, gl, my, wo, gu, ne, xh, ha, nl, yi, haw, no, yo, he, ny, zh, hi, or, zu, hmn, pa, hr, pl
Jina-CLIP-v2 (Koukounas et al., 2024b), Jina-Text-v3 (Sturua et al., 2024)	ar, bn, zh, da, nl, en, fi, fr, ka, de, el, hi, id, it, ja, ko, lv, no, pl, pt, ro, ru, sk, es, sv, th, tr, uk, ur, vi
Multilingual CLIP (Carlsson et al., 2022)- LaBSE ViT-L/14, XLM-R-Large ViT-B/32, XLM-R ViT-L/14, XLM-R-Large ViT-B/16+	af, am, ar, az, bg, bn, bs, ca, cs, cy, da, de, el, en, es, et, fa, fa-AF, fi, fr, gu, ha, he, hi, hr, ht, hu, hy, id, is, it, ja, ka, kk, kn, ko, lt, lv, mk, ml, mn, ms, mt, nl, no, pl, ps, pt, ro, ru, si, sk, sl, so, sq, sr, sv, sw, ta, te, th, tl, tr, uk, ur, uz, vi, zh, zh-TW
M-MPNET, M-MiniLM, MCLIP-ST (Reimers and Gurevych, 2020)	ar, bg, ca, cs, da, de, el, en, es, et, fa, fi, fr, fr-ca, gl, gu, he, hi, hr, hu, hy, id, it, ja, ka, ko, ku, lt, lv, mk, mn, mr, ms, my, nb, nl, pl, pt, pt-br, ro, ru, sk, sl, sq, sr, sv, th, tr, uk, ur, vi, zh-cn, zh-tw, zh

Table 5: List of Multilingual text encoder and multilingual multimodal models and it's supported languages.

for $\lambda = 1$ and $\lambda = 44$) than in Image-to-Text retrieval (0.6% gain in Avg. Recall@10) shown in Table 2.

E Data scaling experiments

Table 8 and 7 show that that our method can learn a strong alignment even with only 1000 English sentences for both I2T and T2I retrieval on XTD dataset. On average across 11 languages, there is insignificant improvement when data is scaled from 50K (89.2% T2I, 89.1% I2T) to 2M sentences (89.3% T2I, 89.5% I2T).

F Image-Text Retrieval: Additional Results

F.1 Language-wise Recall on XM3600 & Multi30K

Tables 9, 10 show language-wise performance of our M2M-aligned models on XM3600 and Multi30K datasets respectively. Interestingly, CLIP $\times$ M-MPNET outperforms Jina-CLIP-v1 $\times$ M-MPNET by 3.7% I2T and 0.5% T2I on Multi30K dataset. Please refer to Table 13 for XTD I2T language-wise breakdown.

F.2 Results on model-supported languages

Similar to our results for Image-Text retrieval in Table 1, in Table 11, we report Recall@10 metric averaged only on languages supported by the respective multilingual text encoder/mutlilingual CLIPs. Supported languages for each model is listed in Table 5.

F.3 Reproducibility experiments

To show that our method's performance is reproducible. We run experiments twice on our method for Image-Text retrieval task, and report mean and standard deviation in Tables 12, 13 to show that the performance is stable across varying random seeds. Rows M1-M6 are defined in Table 1.

G Curation of Synthetic evaluation dataset

For AYA-23-35B, we use translation prompts to generate synthetic data following (Alam et al., 2024). We experiment with zero-shot and 3-shot prompts. We use FLoRes-200 dataset to assess the quality of translation prompts. Zero-shot prompt is fairly straightforward method- we pass the input sentence and prompt the model to generate translation in target language. For 3-shot prompt, for each input english text for which translation has to be generated, we pick 3 examples. These 3 examples are picked from sampling set- created by combining FLoRes-200 validation and test set (excluding current input text). We compute cosine similarity between input text and sampling set using LaBSE, and select top 3 texts and it's corresponding translation of the target language as a few-shot example. The zero-shot translation prompt performs better on the FLoReS-200 dataset (Costa-jussà et al., 2022) across 14 languages[9], achieving a mean spBLEU of 39.7 and mean chrF++ of 51.5, compared to the 3-shot prompt with mean spBLEU of 37.2 and mean chrF++ of 47.4. Given these results, we apply the zero-shot prompt to generate Aya-23-35B translations for all 22 languages. Language-wise spBLEU and chrF++ scores for AYA-23-35B are shown in Table 16, and for backtranslated Indic translations are shown in Table 17. Zero-shot prompt and 3-

[9]ar, zho-Hant, fr, de, he, hi, it, jp, ko, pl, ru, es, tr, vi

Loss	MLP layers	Skip Conn.	Avg.	de	en	es	fr	it	jp	ko	pl	ru	tr	zh
MSE	2	No	88.9	89.9	94.3	90.5	90.1	90.5	82.2	85.3	90.6	86.1	**89.0**	89.7
MSE	2	Yes	88.8	89.7	94.1	90.4	90.0	90.9	81.9	85.2	90.1	86.1	88.9	89.2
$\lambda_1 \cdot \mathcal{L}_{align} + \beta_1 \cdot \mathcal{L}_{str}$	2	No	**89.2**	**90.9**	94.4	91.1	89.5	90.8	**82.4**	85.4	90.6	**87.1**	88.9	**90.1**
$\lambda_1 \cdot \mathcal{L}_{align} + \beta_1 \cdot \mathcal{L}_{str}$	2	Yes	**89.2**	90.5	**94.5**	**91.4**	89.9	91.1	82.3	**85.9**	90.8	86.5	88.1	**90.1**
$\lambda_1 \cdot \mathcal{L}_{align} + \beta_1 \cdot \mathcal{L}_{str}$	4	No	89.1	89.9	94.4	90.8	89.9	91.0	**82.4**	85.6	90.8	86.7	88.5	**90.1**
$\lambda_1 \cdot \mathcal{L}_{align} + \beta_1 \cdot \mathcal{L}_{str}$	1	No	**89.2**	90.4	94.4	90.9	90.3	91.1	**82.4**	85.5	91.0	86.7	88.9	90.0
$\lambda_2 \cdot \mathcal{L}_{align} + \beta_1 \cdot \mathcal{L}_{str}$	2	No	**89.2**	90.4	94.4	90.4	**90.5**	**91.3**	**82.4**	85.7	**91.6**	86.4	88.8	89.2
Similarity Loss	2	No	88.9	90.3	94.4	90.2	90.0	90.4	82.1	85.4	90.6	86.5	88.8	89.3
L1	2	No	86.2	87.2	94.0	88.1	87.4	87.6	78.8	81.0	86.7	83.3	85.9	88.3

Table 6: Comparison of Recall@10 metric across different training losses, and settings- varying number of linear layers, presence or absence of residual connections (Skip Conn.) between linear layers for M2M-aligned Jina-CLIP-v1×M-MPNET on XTD dataset for Text-to-Image retrieval task. $\lambda_1 = 44, \lambda_2 = 1, \beta_1 = 1$.

Scale	Avg.	de	en	es	fr	it	jp	ko	pl	ru	tr	zh
1K	85.8	86.0	92.3	86.3	86.9	86.6	80.3	81.9	88.2	82.8	85.8	86.9
5K	88.5	89.7	93.7	90.2	89.0	90.4	82.5	84.6	90.1	85.2	88.8	89.4
10K	88.8	89.4	94.3	90.3	89.1	91.0	82.1	85.7	90.8	85.4	88.3	89.9
50K	89.2	90.6	94.3	90.9	89.7	91.0	82.3	86.0	**91.0**	86.4	88.9	89.9
100K	89.0	89.7	94.7	90.8	**90.0**	**91.2**	81.9	85.6	90.7	85.8	88.9	89.6
250K	89.2	**90.9**	94.4	**91.1**	89.5	90.8	82.4	85.4	90.6	**87.1**	88.9	90.1
2M	**89.3**	90.5	**94.8**	90.5	**90.0**	91.1	**82.6**	**86.1**	**91.0**	86.4	**89.2**	**90.5**

Table 7: Effect of scaling number of sentences in the training data on the Recall@10 metric for the XTD Text-to-Image Retrieval task using our M2M-aligned Jina-CLIP-v1 × M-MPNET model.

Scale	Avg.	de	en	es	fr	it	jp	ko	pl	ru	tr	zh
1K	84.6	85.4	91.2	85.7	85.0	85.4	78.0	80.3	86.4	81.6	85.4	86.3
5K	88.2	88.4	95.2	89.8	88.7	89.3	81.5	83.5	90.1	85.6	88.7	89.0
10K	88.8	88.9	**95.6**	90.5	89.0	90.5	81.6	84.5	90.9	86.3	90.0	89.5
50K	89.1	89.1	95.1	90.7	89.4	90.5	81.8	85.4	90.9	86.7	90.1	**90.4**
100K	89.0	89.3	95.2	90.7	89.4	90.1	82.2	84.9	90.9	86.5	90.0	90.3
250K	89.3	**89.4**	**95.6**	90.7	89.3	90.6	82.3	85.2	**91.5**	86.5	**90.4**	**90.4**
2M	**89.5**	89.3	95.5	**91.6**	**89.5**	**91.1**	**83.3**	**85.9**	90.8	**87.3**	90.1	**90.4**

Table 8: Effect of scaling number of sentences in the training data on the Recall@10 metric for the XTD Image-to-Text Retrieval task using our M2M-aligned Jina-CLIP-v1 × M-MPNET model.

Retrieval Type	Avg	ar	bn	cs	da	de	el	en	es	fa	fi	fil
T2I	66.4	68.6	31.3	73.2	79.6	81.4	67.4	79.9	75.8	76.2	77.5	10.0
I2T	72.9	77.2	36.3	80.0	86.2	88.1	76.3	85.1	82.3	81.9	84.7	18.0

Retrieval Type	fr	he	hi	hr	hu	id	it	ja	ko	mi	nl	no
T2I	81.5	75.1	60.2	79.3	77.7	85.5	79.3	78.7	72.4	0.7	74.9	79.2
I2T	87.9	83.0	70.6	86.7	83.7	89.9	85.3	85.1	80.9	1.1	80.2	86.2

Retrieval Type	pl	pt	quz	ro	ru	sv	sw	te	th	tr	uk	vi	zh
T2I	76.5	77.3	2.7	80.0	82.3	78.2	4.5	29.1	79.1	74.7	76.5	81.8	80.9
I2T	83.3	83.4	6.4	87.6	88.4	85.0	9.3	38.8	85.6	81.5	83.7	88.9	86.9

Table 9: Recall@10 across 36 languages for XM3600 on I2T and T2I retrieval task using M2M-aligned Jina-CLIP-v1 × M-MPNET.

Model	T2I					I2T				
	Avg	cs	de	en	fr	Avg	cs	de	en	fr
Jina-CLIP-v1 × M-MPNET	90.1	88.2	**89.6**	92.4	90.3	89.7	87.3	89.4	92.0	90.1
CLIP × M-MPNET	**90.6**	**88.4**	89.0	**93.4**	**91.4**	**92.0**	**89.8**	**91.2**	**94.9**	**92.3**

Table 10: Recall@10 across 4 languages for Multi30K on I2T and T2I retrieval task using M2M-aligned Jina-CLIP-v1 × M-MPNET.

Models	XM3600		Multi30K	
	T2I	I2T	T2I	I2T
English-only Zero-shot Baseline Models				
E1: CLIP (ViT-L 336px)	77.3	87.1	93.5	**95.8**
E2: Jina-CLIP-v1	85.7	91.8	93.5	93.6
E3: K-ALIGN	**87.0**	**92.0**	**95.9**	**95.8**
Multilingual Multimodal Models Trained on Supervised Multimodal and/or Multilingual Data				
T2: MCLIP-ST	57.6	71.1	80.7	83.4
T5: LABSE ViT-L/14	77.0	87.5	90.9	93.7
T6: XLM-R-L ViT-B/32	79.6	89.0	89.2	91.0
T7: XLM-R ViT-L/14	80.9	89.6	92.2	94.4
T8: XLM-R-L ViT-B/16+	86.5	91.9	93.9	94.2
T9: Jina-CLIP-v2	**90.1**	**93.9**	**94.3**	**94.5**
M2M-aligned Multilingual Multimodal models Trained on only English Text data				
M1: Jina-CLIP-v1 × LaBSE	64.8	67.6	78.7	75.2
M2: Jina-CLIP-v1 × M-MiniLM	68.6	75.5	87.8	85.7
M3: Jina-CLIP-v1 × JinaTextV3	75.1	80.2	89.0	88.3
M4: Jina-CLIP-v1 × M-MPNET	**76.9**	**83.8**	90.0	89.7
M5: CLIP × M-MPNET	64.6	77.1	**90.6**	**92.0**
M6: K-ALIGN × M-MPNET	68.7	78.7	90.4	90.0

Table 11: Performance of M2M-align models in comparison with English and Mutlingual CLIP-like models on Recall@10 metric for supported languages for XM3600 and Multi30K datasets.

Models	Avg.	de	en	es	fr	it	jp	ko	pl	ru	tr	zh
M1	82.6±0.3	82.8±0.4	86.7±0.4	83.8±0.1	84.6±0.2	84.7±0.2	76.5±0.4	80.5±0.6	84.8±0.4	80.2±0.2	81.1±0.6	83.3±0.4
M2	86.5±0.0	87.4±0.1	93.9±0.3	88.5±0.4	87.9±0.1	87.4±0.1	80.7±0.1	75.0±0.2	89.2±0.1	84.9±0.1	86.0±0.3	90.6±0.6
M3	88.0±0.2	91.0±0.0	95.4±0.1	90.1±0.8	90.5±0.6	91.2±0.1	80.0±0.6	80.3±0.3	90.2±0.1	85.9±0.4	87.8±0.6	85.2±0.4
M4	89.2±0.0	90.7±0.4	94.5±0.1	90.9±0.4	89.9±0.6	90.9±0.1	82.2±0.4	85.8±0.5	90.9±0.4	86.7±0.6	88.8±0.2	90.1±0.0
M5	84.3±0.1	85.1±0.4	91.1±0.1	85.8±0.2	85.4±0.4	86.1±0.4	77.9±0.1	80.4±0.4	84.8±0.1	81.7±0.1	84.7±0.2	84.2±0.4
M6	86.8±0.0	87.3±0.4	92.9±0.2	89.7±0.1	87.9±0.1	88.4±0.1	79.0±0.4	83.1±0.1	88.7±0.1	83.3±0.1	86.9±0.1	87.7±0.2

Table 12: Performance of M2M-aligned Jina-CLIP-v1 × M-MPNET on Recall@10 metrics averaged (± standard deviation) over 2 different runs across 11 languages for Text-to-Image retrieval task on XTD dataset.

Models	Avg.	de	en	es	fr	it	jp	ko	pl	ru	tr	zh
M1	80.1±0.0	79.6±0.3	87.3±0.0	81.0±0.1	81.8±0.1	83.3±0.3	72.1±0.1	75.8±0.5	82.7±0.0	77.6±0.3	79.5±0.1	80.7±0.4
M2	84.9±0.1	85.1±0.1	94.0±0.2	86.2±0.3	85.7±0.1	86.9±0.4	79.7±0.1	69.9±0.9	88.2±0.1	84.0±0.1	85.6±0.2	88.7±0.3
M3	87.6±0.1	90.6±0.0	94.7±0.1	90.2±0.2	89.5±0.1	90.2±0.2	79.9±0.5	80.4±0.1	88.1±0.1	84.9±0.4	87.3±0.2	87.3±0.3
M4	89.3±0.0	89.3±0.1	95.8±0.2	90.8±0.1	89.5±0.2	90.6±0.0	82.6±0.4	85.4±0.2	91.4±0.2	86.4±0.1	90.2±0.4	90.5±0.1
M5	85.8±0.1	85.9±0.2	93.8±0.1	89.0±0.0	87.2±0.5	88.7±0.4	76.8±0.1	82.0±0.4	86.8±0.1	82.6±0.1	85.3±0.1	86.4±0.1
M6	86.1±0.1	86.3±0.3	94.6±0.5	88.3±0.4	87.6±0.3	89.0±0.1	78.1±0.1	82.2±0.1	85.9±0.2	82.3±0.5	84.8±0.3	87.3±0.0

Table 13: Performance of M2M-aligned Jina-CLIP-v1 × M-MPNET on Recall@10 metrics averaged (± standard deviation) over 2 different runs across 11 languages for Image-to-Text retrieval task on XTD dataset.

shot prompt templates are listed in Table 14 and 15.

> You are an expert in translations. Your task is to accurately translate the following text into [target language].
> Input text: [input test sentence]
> Translation:

Table 14: Zero-shot prompt used generating translation from AYA-23-35B. Text in square bracket is a placeholder for actual input

> You are an expert in translations. Your task is to accurately translate the following text into [target language].
>
> Here are a few examples to help you understand the format:
>
> Example 1:
> Input text: [input text 1]
> Translation: [translation 1]
>
> Example 2:
> Input text: [input text 2]
> Translation: [translation 2]
>
> Example 3:
> Input text: [input text 3]
> Translation: [translation 3]
>
> Now, translate the following text:
>
> Input text: [input test sentence]
> Translation:

Table 15: 3-shot prompt template used to compare effect of few-shots on translation quality for AYA-23-35B. Text in square bracket is a placeholder for actual input.

H CLAP

H.1 Language-wise Recall on Synthetic Evaluation Dataset.

We show language-wise performance of M2M-aligned CLAP-general × M-MPNET on AudioCaps in Table 18 and Clotho in Table 19.

H.2 Quantifying the qualitative analysis and more examples

We see in Table 3 that M2M-aligned models don't match the performance of baseline CLAP models. For English, qualitative analysis revealed that the retrieved audio for a query text had high semantic similarity. To verify our qualitative analysis, we perform following quantitative test. For each query text, we retrieve top five audios using M2M-aligned model CLAP-General × M-MPNET. Next, we compute cosine similarity between query text and captions of retrieved audio using CLAP-general model. On average, we see higher cosine similarity for CLAP-general × M-MPNET (0.7) compared to CLAP-general (0.65), demonstrating semantic agreement between CLAP-general and retrieved audio. More examples are listed in Table 20.

I Cross-lingual Text-to-Image Generation.

Both Inception score and FID scores are computed using torch-fidelity (Obukhov et al., 2020) package [10]. Language-wise FID scores shown in Table 21. For English, our aligned model gives better FID score than FLUX-CLIP though both are still high compared to FLUX (upper-bound/skyline model). More examples of generated images are shown in Figure 4, Figure 5 & Figure 6.

[10] https://github.com/toshas/torch-fidelity

Prompts	Avg.	ar	zh	fr	de	he	hi	it	jp	ko	pl	ru	es	tr	vi
spBLEU															
3-shot prompt	37.2	**46.7**	20.6	67.3	50.0	37.2	8.9	**64.8**	25.4	16.7	22.6	**56.5**	47.5	**36.0**	19.9
zero-shot prompt	**39.7**	21.6	**21.3**	**69.2**	**56.2**	**55.6**	**28.2**	54.2	**28.6**	**17.0**	**33.7**	51.6	**51.9**	29.1	**37.4**
chrF++															
3-shot prompt	47.4	**36.5**	**31.4**	63.0	55.8	71.8	18.6	**80.3**	29.4	17.7	39.0	**61.1**	55.2	60.0	44.0
zero-shot prompt	**51.5**	29.0	26.3	**64.5**	**58.6**	**77.6**	**49.0**	78.0	**31.4**	**22.3**	**41.0**	58.5	**57.8**	**63.8**	**62.7**

Table 16: spBLEU and chrF++ scores for zero-shot and 3-shot prompts for FLoRes-200 using AYA-23-35B model. zh in the table denotes Chinese Traditional (zh-Hant).

Test-Dataset	Avg.	bn	gu	hi	kn	ml	mr	ne	pa	ta	te	ur
spBLEU												
AudioCaps	48.7	38.3	24.3	39.3	56.2	79.5	19.1	33.0	53.1	64.3	100.0	28.1
CLOTHO	47.4	46.3	51.4	51.4	31.5	28.7	67.9	51.4	48.1	51.4	43.3	50.4
chrF++												
AudioCaps	63.6	60.6	54.0	58.6	73.0	74.2	37.7	46.9	82.1	61.5	100.0	51.5
CLOTHO	59.6	43.8	64.0	65.9	46.0	52.4	68.0	60.5	65.3	65.9	58.2	65.2

Table 17: spBLEU and chrF++ scores on English backtranslations of AudioCaps and Clotho dataset using IndicTrans2 models.

Retrieval Type	Avg	ar	bn	cs	de	el	en	fr	gu	he	hi	id	it
T2A	48.3	45.7	23.2	57.4	58.2	55.2	77.2	60.5	38	48.5	52.6	58.6	57.9
A2T	60.8	61	35.2	65.8	68.8	68.9	81.5	69.5	53.1	61.8	63.2	68.3	67.3

Retrieval Type	ja	kn	ko	ml	mr	nl	ne	pa	fa	pl	pt	ro	ru
T2A	50.2	23.7	48.1	26.2	46.4	61	33.9	22.4	54.6	52	61.1	58.4	49.6
A2T	66.7	38.2	63.1	43.2	61.9	69.9	47.4	34.8	69.6	65.9	70.5	68	64.8

Retrieval Type	es	ta	te	tr	uk	ur	vi	zh (hans)	zh (hant)
T2A	59.5	29.4	25.4	56.3	46.1	44.4	56.4	53.1	50.8
A2T	67.5	42.9	38.7	63.8	62.1	57.3	67.2	69.7	67.9

Table 18: Recall@10 metric across 34 languages on AudioCaps dataset for Audio-to-Text (A2T) and Text-to-Audio (T2A) retrieval task using M2M-aligned CLAP-general $\times$ M-MPNET model.

Retrieval Type	Avg	ar	bn	cs	de	el	en	fr	gu	he	hi	id	it
T2A	33.3	34.3	18.8	37.5	37.7	35.9	47.8	40.4	25.9	34.1	36.8	39.8	38.2
A2T	39.6	42	24.1	41.5	43.2	41.2	50.6	44.6	34.8	38.9	42.5	44.8	44.1

Retrieval Type	ja	kn	ko	ml	mr	nl	ne	pa	fa	pl	pt	ro	ru
T2A	38.6	17.2	35	20.8	29.3	38.7	23.9	17.9	37	36.6	39.4	38.3	35.5
A2T	45.5	25.6	42.8	27.7	36.9	44.8	31	24.9	43.5	42.4	44.3	43.4	44.7

Retrieval Type	es	ta	te	tr	uk	ur	vi	zh (hans)	zh (hant)
T2A	39.7	20.9	17.8	36.5	33.7	31.3	39	40	39.8
A2T	45.6	27.9	24.6	42.7	42	37.5	45	45.3	44.9

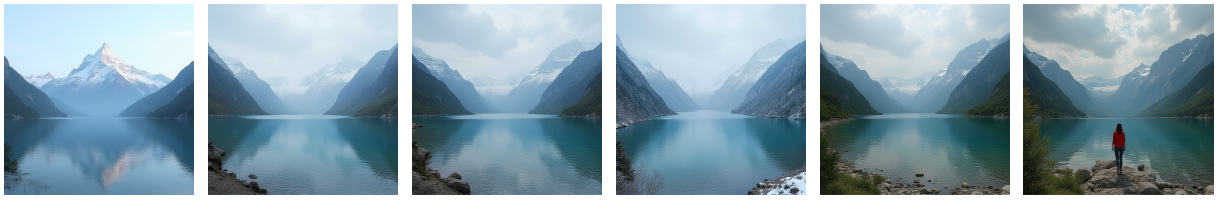
Table 19: Recall@10 metric across 34 languages on Clotho dataset for Audio-to-Text (A2T) and Text-to-Audio (T2A) retrieval task using M2M-aligned CLAP-general $\times$ M-MPNET model.

Query Text	Captions of Retrieved Audios Rank 1	Rank 2	Rank 3	Ground Truth (Rank 9)
Water flows and people speak in the distance	Water splashing with multiple voices in background	Water is trickling, and a man talks	A river stream flowing followed by a kid talking	Running water and distant speech
	A man shouting as a stream of water splashes and a crowd of people talk in the background	Splashing water and quiet murmuring	A large volume of water is rushing, splashing and gurgling, and an adult male speaks briefly	A stream of water rushing as a man shouts in the distance
	A plastic clack followed by a man talking as a stream of water rushes and a crowd of people talk in the background	Bubbles gurgling and water spraying as a man speaks softly while crowd of people talk in the background	A stream of water rushing and trickling followed by a young man whooshing	Water rushing loudly while a man yells in the background
	Water splashes and a man speaks	Water trickling and faint, muffled speech	Sounds of a river with man briefly mumbling	A large volume of water is rushing fast, splashing and roaring, and an adult male shout in the background
	Water is falling, splashing and gurgling, a crowd of people talk in the background, and an adult male speaks in the foreground	Water spraying and gurgling as a man speaks and a crowd of people talk in the background	A stream burbles while a man speaks	Water flows and people speak in the distance
Query Text	Rank 1	Rank 2	Rank 3	Ground Truth (Rank 10)
A frog croaks with speech and thumping noises in the background	Frogs croaking together with a man speaking followed by rustling	Frogs croaking and a humming with insects vocalizing	Frogs croaking with rustling in the background	Nature sounds with a frog croaking
	A man talking followed by plastic clunking and rattling as frogs croak and crickets chirp	A frog croaking and insects vocalizing with a humming	Two instances of bird wings flapping while frogs are croaking	A frog chirping as a woman talks over an intercom and water splashes in the background followed by wood falling on a hard surface
	A man talking followed by plastic creaking and clacking as frogs croak and crickets chirp	A croaking frog with brief bird chirps	A group of frogs croaking as plastic flutters in the background	A frog chirping with distant speaking of a person
	Several frogs chirping near and far with men speaking and some banging	Crickets chirping very loudly	Frogs chirp loudly	A frog croaking as a woman talks through an intercom while water is splashing and wood clanks in the background
	A man speaking as frogs croak and crickets chirp while a motorboat engine runs alongside several plastic clacks and clanging	Ambient horror music plays as birds chirp and frogs croak	High pitched croaking of frogs with some rustling	A frog croaks with speech and thumping noises in the background

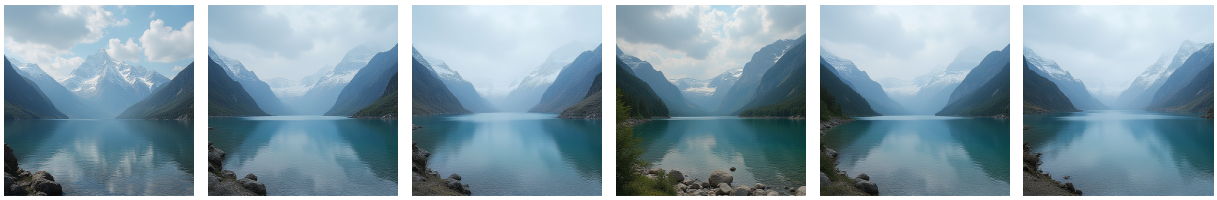
Table 20: Captions of Audio retrieved for a Query text (Text-to-Audio retrieval task) using M2M-aligned CLAP-General $\times$ M-MPNET

Models	FID-30K (↓)									
	en	fr	el	he	id	ko	fa	ru	es	hi
FLUX	23.4	-	-	-	-	-	-	-	-	-
FLUX-T5	23.4	-	-	-	-	-	-	-	-	-
FLUX-CLIP	40.9	-	-	-	-	-	-	-	-	-
FLUX × M-MPNET	36.9	41.8	46.6	46.9	40.0	45.4	43.0	47.2	41.1	45.1

Table 21: FID scores computed on our MSCOCO 30K synthetic multilingual evaluation dataset.



(a) FLUX (en) (b) Ours (en) (c) Ours (el) (d) Ours (fa) (e) Ours (fr) (f) Ours (he)



(g) FLUX CLIP (en) (h) Ours (ru) (i) Ours (hi) (j) Ours (id) (k) Ours (ko) (l) Ours (es)

Figure 4: Images generated by FLUX text-to-image model using the prompt "a snow caped mountain is behind a large lake" in multiple languages. Our M2M-aligned model produces similar quality images compared to baseline FLUX (both T5 and CLIP encoders), and FLUX-CLIP models.

(a) FLUX (en) (b) Ours (en) (c) Ours (el) (d) Ours (fa) (e) Ours (fr) (f) Ours (he)

(g) FLUX CLIP (en) (h) Ours (ru) (i) Ours (hi) (j) Ours (id) (k) Ours (ko) (l) Ours (es)

Figure 5: Images generated by FLUX text-to-image model using the prompt "Assortment of colorful flowers in glass vase on table." in multiple languages. Our M2M-aligned model produces similar quality images compared to baseline FLUX (both T5 and CLIP encoders), and FLUX-CLIP models.

(1) T5 prompt: “A photo of: ”

FLUX (en) Ours (en) Ours (el) Ours (fa) Ours (fr) Ours (he)

FLUX CLIP (en) Ours (ru) Ours (hi) Ours (id) Ours (ko) Ours (es)

(2) T5 prompt: “add a book: ”

FLUX-T5 (en) Ours (en) Ours (el) Ours (fa) Ours (fr) Ours (he)

FLUX CLIP (en) Ours (ru) Ours (hi) Ours (id) Ours (ko) Ours (es)

(3) T5 prompt: “add a book on bed: ”

FLUX-T5 (en) Ours (en) Ours (el) Ours (fa) Ours (fr) Ours (he)

FLUX CLIP (en) Ours (ru) Ours (hi) Ours (id) Ours (ko) Ours (es)

Figure 6: Images generated by FLUX models using the prompt “A cat sitting on a bed behind a book” in multiple languages. Our M2M-aligned model produces similar images but with missing objects (book, bed) compared to FLUX models (T5 and CLIP encoders). T5 prompts help mitigate this issue, as shown in sub-figures (2) & (3).