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

Piyush Singh Pasi
Amazon*
piyush.singh.pasi@gmail.com

Abstract

Multimodal models excel in English, supported by abundant image–text and audio–text data, but performance drops sharply for other languages due to limited multilingual multimodal resources. Existing solutions rely on machine translation, while advances in multilingual text modeling remain underutilized. We introduce M2M, a lightweight alignment method that learns only a few linear layers—using English text alone—to map multilingual text embeddings into multimodal space. Despite its simplicity, M2M matches baseline performance in English (94.9% Recall@10) and achieves strong zero-shot transfer (89.5% Recall@10 averaged across 11 languages, 10 unseen) on XTD Text-to-Image retrieval. Qualitative t-SNE visualizations show that multilingual embeddings align tightly with multimodal representations, while weight analysis reveals that the transformation reshapes embedding geometry rather than performing trivial rotations. Beyond image–text retrieval, M2M demonstrates robustness across datasets and tasks, extending to Audio–Text retrieval and Text-to-Image generation. We release code and checkpoints[1] along with multilingual evaluation datasets: MSCOCO Multilingual 30K[2], AudioCaps Multilingual[3], and Clotho Multilingual[4].

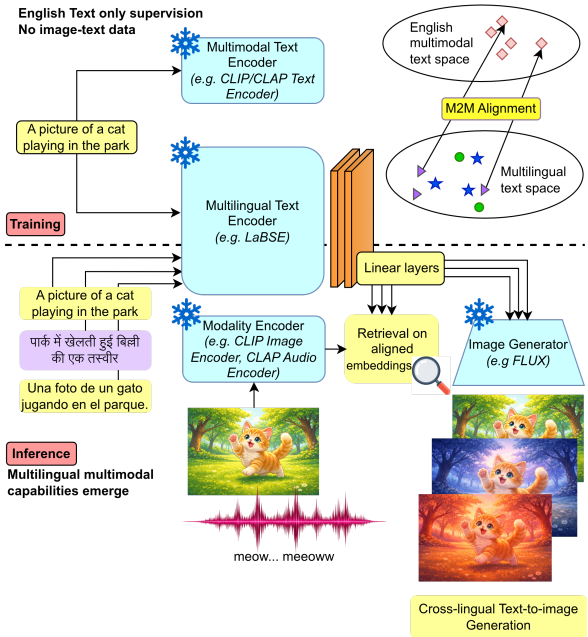


Figure 1: Overview of M2M. Using only English text supervision, we learn a lightweight linear mapping that aligns multilingual text embeddings to a frozen multimodal text space (e.g., CLIP). English acts as a shared anchor during training, aligning multilingual text representations (triangles) to the multimodal text space (diamonds). This alignment implicitly transfers to other languages (stars and circles) without requiring any additional multilingual or multimodal supervision.

1 Introduction

Humans naturally align information across modalities, associating visual objects/sounds with words. For example, once a person has learned to associate the visual concept of a *cat* with the English word "cat", learning that the Spanish word "gato" refers to the same concept allows the object–word association for "gato" to emerge implicitly, without requiring visual supervision in Spanish.

Existing multimodal models, such as CLIP (Radford et al., 2021) and CLAP (Elizalde et al., 2023), are primarily trained on large-scale English multimodal corpora and rely on explicit multimodal supervision. Extending these models to additional languages typically requires substantial multilingual image–text or audio–text data, either by training from scratch or fine-tuning pretrained models (Carlsson et al., 2022; Yan et al., 2024; Koukounas et al., 2024b). Acquiring such multilingual multimodal resources is expensive or infeasible, particularly for low-resource languages. By contrast, multilingual text encoders have demon-

*Work done outside of Amazon, not related to role diretly
[1]GitHub: piyushsinghpasi/M2M
[2]HF: piyushsinghpasi/mscoco-multilingual-30k
[3]HF: piyushsinghpasi/audiocaps-multilingual
[4]HF: piyushsinghpasi/clotho-multilingual

Findings of the Association for Computational Linguistics: EACL 2026, pages 2750–2771
March 24-29, 2026

strated strong cross-lingual generalization using only large-scale text corpora and self-supervised objectives (Devlin et al., 2019; Radford, 2018), but this capability remains largely disconnected from pretrained multimodal representations.

In this work, we propose M2M, a simple, data-efficient, and parameter-efficient approach to bridge multilingual text and multimodal latent spaces using English text as a shared anchor. Inspired by how humans learn, our method does not require explicit multimodal signals for each language—English textual data alone is sufficient for alignment. We learn a lightweight projection map, implemented as a few linear layers, while keeping all pretrained encoders frozen and training with MSE and structure-preserving losses. Despite its simplicity, this alignment enables multilingual text representations to participate directly in multimodal tasks, including retrieval and generation, without observing any multilingual image–text or audio–text pairs. Rather than aiming to replace large-scale multilingual multimodal pretraining, our goal is to show that improved latent-space alignment can recover much of the same capability at a fraction of the data and computational cost.

Prior work has shown the effectiveness of linear projection maps for aligning latent spaces using English multimodal data, mostly in classification settings (Maiorca et al., 2024b; Rosenfeld et al., 2022). We extend this approach to a broader setting: multilingual and multimodal alignment across retrieval and generative tasks. Our results show that strong multilingual multimodal behavior can emerge from lightweight alignment alone when robust multilingual text encoders are used. To summarize, our contributions are as follows:

1. We propose M2M, a lightweight alignment method that maps multilingual text representations into pretrained multimodal latent spaces using only monolingual (English) text data. Despite its simplicity, this approach enables cross-task, cross-modality transfer, allowing multilingual text representations to participate in retrieval and generation tasks without observing any multilingual multimodal data during training.

2. M2M is highly data-efficient, achieving strong performance with as few as $\sim$1K sentences, and parameter-light, requiring only a few linear layers ($\sim$1–2M parameters). It generalizes across architectures, modalities (image and audio), tasks (Image–Text and Audio–Text retrieval, and Text-to-Image generation), and languages—including those unseen during multimodal pretraining.

3. We construct synthetic multilingual evaluation benchmarks for multimodal retrieval and generation: (i) Audio–Text retrieval datasets in 33 languages derived from AudioCaps (Kim et al., 2019) (160K samples) and Clotho (Drossos et al., 2019) (172K samples), and (ii) MSCOCO-30K captions translated into 9 additional languages (270K samples) for cross-lingual Text-to-Image generation. These datasets provide a unified and reproducible benchmark for evaluating multilingual multimodal 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.

In contrast, our method targets multilingual alignment without relying on multilingual multimodal supervision (e.g., image–text pairs) or encoder fine-tuning. Using simple and intuitive training losses, a small number of linear projection layers, and only English text data, we demonstrate strong multilingual transfer across a broad range of multimodal tasks, including image–text and audio–text retrieval, as well as text-to-image generation.

Latent Space Translation. Latent space translation aims to map representations between distinct latent spaces in order to enable information sharing across modalities or languages. Prior work broadly follows two directions: (i) aligning spaces using relative representations (Moschella et al., 2022; Norelli et al., 2022), and (ii) learning direct transformation mappings between source and target spaces (Gower, 1975; Maiorca et al., 2024b; Lähner and Moeller, 2024). These techniques have been successfully applied to tasks such as cross-modal classification and generative modeling. A more recent extension is the Inverse Relative Projection method (Maiorca et al., 2024a), which converts source representations into a relative form before mapping them to a target space, enabling the translation of monolingual text representations into multilingual ones.

Building on this line of work, our approach learns a linear mapping between multilingual and multimodal latent spaces. By leveraging English text as a shared anchor between these spaces, we enable multilingual multimodal transfer without requiring multilingual multimodal training data.

3 Methodology

Our method, M2M, is a simple alignment approach that learns a small projection network to align multilingual latent spaces with multimodal latent spaces using English text representations. While we focus on dual-modality multimodal models, the method naturally extends to more than two modalities.

Consider an English (monolingual) multimodal model $\mathcal{M}_e = (T_e, X_e)$ for language e, where T_e is the text encoder and X_e represents any other modality encoder (e.g., image, audio). We assume representations from 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. For a sentence s in language e, let $z_e = T_e(s)$ denote its multimodal representation and $z_m = T_m(s)$ its multilingual representation, with $z_e \in \mathbb{R}^{d_e}$ and $z_m \in \mathbb{R}^{d_m}$. Since both text encoders represent the same sentence s, in an *all-aligned* world, z_e and z_m would be identical. In practice, however, they differ due to distinct objectives and training data.

Our goal is therefore to align the multilingual and multimodal latent spaces using English—the common language between these spaces—as an anchor. This alignment enables multimodal tasks on non-English languages, for which no multimodal training data is ever seen; any downstream performance on these languages arises purely from the learned alignment.

To achieve this, we learn a projection map $\mathcal{F} : \mathbb{R}^{d_m} \to \mathbb{R}^{d_e}$ that transforms z_m into z_e, using text-only alignment data in language e (English). The alignment data must be semantically consistent with the downstream task (e.g., image captions for image–text retrieval, audio captions for audio–text retrieval).

The projection map $\mathcal{F}$—implemented as a few linear layers—is the only learned component, while all encoders (T_e, T_m, X_e) remain frozen. During inference, multilingual text is encoded by T_m, mapped via $\mathcal{F}$, and then directly compared with X_e, producing task-compatible representations for retrieval or generation. In this setup, z_e serves as an anchor guiding the translation of multilingual embeddings into the multimodal space. We use mean squared error (MSE) as our primary loss function.

$$z_{m\to e} = \mathcal{F}(z_m) \tag{1}$$

$$\mathcal{L}_{\text{align}} = \text{MSE}(z_{m\to e}, z_e) \tag{2}$$

To derive additional supervision, we enforce structure preservation within each batch B. Let $\{z_e^i\}_{i=1}^{|B|}$ and $\{z_{m\to e}^i\}_{i=1}^{|B|}$ denote the target multimodal embeddings and their projected multilingual counterparts. We compute pairwise cosine similarities:

$$R_e = \text{cos\_sim}(z_e^i, z_e^j)_{i,j=1}^{|B|}, \tag{3}$$

$$R_{m\to e} = \text{cos\_sim}(z_{m\to e}^i, z_{m\to e}^j)_{i,j=1}^{|B|}, \tag{4}$$

where $R_e, R_{m\to e} \in \mathbb{R}^{|B|\times|B|}$. Let $\text{triu}(\cdot)$ denote the upper-triangular part of a matrix, excluding the diagonal. The structure-preserving loss is then

$$\mathcal{L}_{\text{str}} = \text{MSE}\big(\text{triu}(R_e), \text{triu}(R_{m\to e})\big). \tag{5}$$

The final objective combines the alignment and structure-preserving terms:

$$\mathcal{L} = \lambda\, \mathcal{L}_{\text{align}} + \beta\, \mathcal{L}_{\text{str}}. \tag{6}$$

We experimented with alternative objectives such as L1 loss and similarity loss ($1 - \text{cosine}(z_e, z_{m\to e})$), but these underperform compared to $\mathcal{L}$. MSE is particularly effective because it encourages $z_{m\to e}$ to fully substitute for z_e in the latent space, rather than focusing solely on angular alignment as in contrastive or cosine-based losses. We avoid token- or word-level alignment, which overemphasizes linguistic form over semantics, and do not introduce a reverse mapping $\mathcal{F}_{e\to m}$ since it would disrupt the existing alignment between the other modality encoder X_e and T_e. For

Loss	Linear layers	Skip Conn.	T2I	I2T
V1: MSE	2	Yes	88.9	88.6
V2: MSE	2	No	89.0	88.7
V3: Similarity Loss	2	No	88.8	88.7
V4: L1	2	No	86.7	83.9
V5: Ours (eq 6)	2	Yes	89.2	**89.4**
V6: Ours (eq 6)	2	No	**89.5**	**89.4**
V7: Ours (eq 6)	4	No	89.1	89.2
V8: Ours (eq 6)	1	No	89.3	89.3

Table 1: Comparison of I2T and T2I Recall@10 (averaged across 11 languages) for different training losses, linear layers, and residual connections (Skip Conn.) with M2M-aligned Jina-CLIP-v1×M-MPNET on XTD test dataset. $\lambda = 48, \beta = 1$.

retrieval tasks, both $\mathcal{L}_{\text{align}}$ and $\mathcal{L}_{\text{str}}$ are computed on L2-normalized embeddings to optimize for downstream evaluation which uses cosine similarity metric. For generative tasks (e.g., text-to-image generation), we omit normalization and L_{str} (see Section 7).

4 Exploring Alignment Design Space

We investigate the impact of varying number of linear layers (1, 2, 4), adding or removing residual connections (He et al., 2015) in $\mathcal{F}$, 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 $\mathcal{F}$. 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.

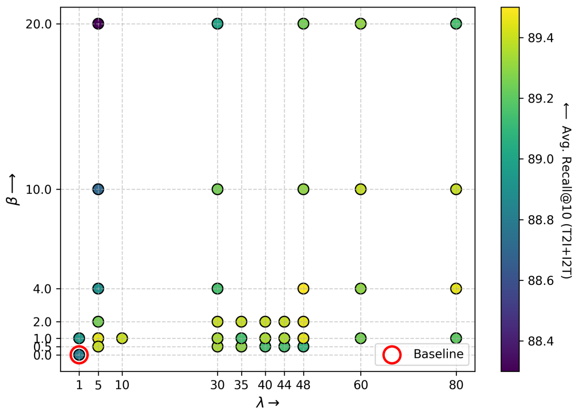


Figure 2: Impact of λ and β on XTD image–text retrieval. Increasing λ while reducing β leads to consistent performance gains.

Results and discussion. Table 1 shows that our proposed loss (eq. 6) consistently outperforms alternatives. Two linear layers and no skip connection (row V6) achieves the best performance, yielding absolute gains of up to 0.7% over MSE (row V2) and similarity loss (row V3), and around 3–5% over L1 loss (row V4). Varying linear layers or using a skip connection has minor effects (rows V6-V8), indicating the model is robust to these architectural choices. Assigning higher weight to $\mathcal{L}_{\text{align}}$ ($\lambda = 48$) versus $\mathcal{L}_{\text{str}}$ ($\beta = 1$) yields a 0.5% gain compared to equal weighting ($\lambda = 1, \beta = 1$), as shown in Figure 2. Overall, the combination of two linear layers (~1M parameters) without a skip connection and $\lambda = 48, \beta = 1$ provides the strongest results. See appendix D for detailed numbers.

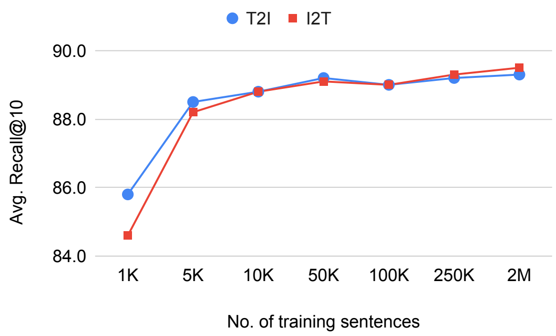


Figure 3: Effect of scaling training data on XTD eval set for M2M-aligned model Jina-CLIP-v1 × M-MPNET. Recall@10 averaged across all languages in XTD.

Data scaling experiments (Figure 3) using the optimal configuration (2 linear layers, no residuals, $\lambda = 48, \beta = 1$) show that M2M achieves 85.8% Avg. Recall@10 with just 1,000 English sentences, without any multilingual or multimodal data. Performance saturates beyond 250K sentences; scaling to 2M sentences provides minimal improvements (0.1–0.2%).

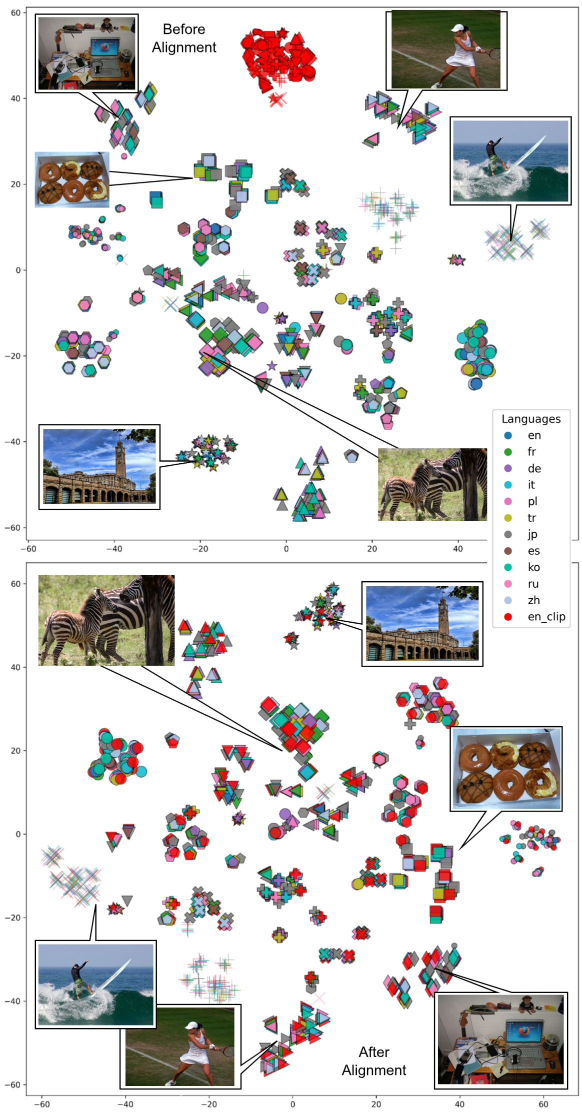


Figure 4: t-SNE visualization (perplexity = 32) of text embeddings before and after alignment. Marker shapes denote visual clusters and colors indicate languages, with English Jina-CLIP-v1 text embeddings (z_e or en_clip) in red. Before alignment (top), text embeddings (z_m & z_e) are fragmented; after alignment (bottom), multilingual captions ($z_{m\rightarrow e}$) and Jina-CLIP-v1 text embeddings (z_e) align closely with shared visual clusters.

Alignment quality. For the best configuration (row V6 in Table 1), we visualize alignment using t-SNE (Figure 4). To avoid language bias, we first cluster image embeddings of Jina-CLIP-v1 (J-CLIP) from the XTD test set using KMeans (K=100), then select 17 clusters via farthest-cluster sampling from the 50 largest clusters, excluding very small clusters. From each cluster, we sample up to 10 points to prevent overcrowding. We compute t-SNE jointly for text embeddings from J-CLIP (z_e) and M-MPNET (z_m) before alignment (Figure 4, top), and for J-CLIP embeddings (z_e) with aligned embeddings ($z_{m\rightarrow e}$) after alignment (Figure 4, bottom) to visualize the effect of M2M. As shown, J-CLIP and M-MPNET embeddings occupy distinct regions before alignment; after alignment, J-CLIP embeddings align with the multilingual embeddings across clusters, demonstrating effective cross-lingual alignment.

Weight analysis. Using our best configuration (row V6 in Table 1), we analyze the effective linear map $W_{eff} = W_2 W_1$. The singular value spectrum revealed that the map focuses on a compact, semantically relevant subspace, with an effective rank of ~ 204. The orthogonality deviation ($\|W^\top W - I\|_F \approx 554$) indicates that the transformation involves substantial *mixing and rescaling* rather than a simple rotation. The effective bias is small ($\|\mathbf{b}_{\text{eff}}\| \approx 1.3$), suggesting that alignment primarily reshapes the geometry of embeddings rather than just shifting them. Because the effective rank is lower than $d_e = d_m = 768$, Jina-CLIP-v1 text embeddings likely capture language-specific features that are absent in the language-agnostic M-MPNET space. Pairwise cosine distances between nearly identical sentences further support this hypothesis: Jina-CLIP-v1 embeddings vary the most (0.03–0.08), multilingual MPNET embeddings are tighter (0.01–0.04), and the mapped embeddings fall in between (0.01–0.05). This confirms that Jina-CLIP-v1 embeddings encode language-specific style variations, whereas the mapped embeddings preserve semantic consistency across sentence variants. See additional details and box plots in Appendix E.

5 Image-Text Retrieval

Experimental setup. We evaluate English vision–language models ($\mathcal{M}_e$): CLIP (Radford et al., 2021), Jina-CLIP-v1 (Koukounas et al., 2024a), K-ALIGN (Yoon et al., 2022), & SigLIP[5] (Zhai et al., 2023), and multilingual text encoders (T_m):

[5]HF: google/siglip-so400m-patch14-384

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**	–	–	–	–	–	–
T11: SigLIP	67.2	87.9	96.7	93.3	91.0	90.2	19.7	25.8	75.5	69.3	63.3	26.2	71.6	**98.3**	40.1	51.9	85.4	87.7
T12: SigLIP2	92.6	**94.6**	96.7	**95.8**	95.0	**96.1**	80.2	91.3	**95.8**	**92.1**	91.7	89.1	93.7	97.9	74.6	81.5	**96.2**	96.2
T13: LLM2CLIP	92.1	92.0	**97.3**	93.8	92.1	93.6	91.9	88.6	93.2	89.4	87.7	93.8	**94.0**	97.8	72.8	83.5	96.0	**96.7**
M2M-aligned Multilingual Multimodal models using English-only Text data																		
M1: Jina-CLIP-v1 × LaBSE	82.7	82.4	86.5	83.6	84.8	85.0	76.5	80.3	85.4	80.7	81.4	82.6	80.0	86.8	62.9	65.6	79.0	75.7
M2: Jina-CLIP-v1 × M-MiniLM	86.4	86.7	93.8	88.1	88.4	87.6	80.5	74.8	89.0	85.2	86.3	**90.2**	84.9	93.7	57.7	64.5	88.0	85.9
M3: Jina-CLIP-v1 × JinaTextV3	88.0	**91.1**	**94.9**	89.6	**90.6**	90.9	80.2	80.9	90.4	85.5	88.0	85.9	87.5	95.0	**67.2**	72.2	87.8	87.5
M4: Jina-CLIP-v1 × M-MPNET	**89.5**	90.5	94.7	**91.9**	90.5	**91.1**	**82.4**	**85.8**	**91.2**	**86.8**	**89.2**	89.9	**89.4**	**95.2**	66.3	**73.1**	89.9	89.7
M5: CLIP × M-MPNET	84.6	85.6	90.8	86.6	85.3	86.2	79.1	80.5	85.1	82.3	84.6	84.5	86.2	93.9	56.2	67.3	90.1	**92.1**
M6: K-ALIGN × M-MPNET	86.8	87.5	92.6	89.6	87.9	87.8	78.9	83.0	89.0	83.5	87.0	87.9	86.0	94.4	59.0	68.2	91.0	90.2
M7: SigLIP × M-MPNET	86.1	87.2	92.3	86.7	87.2	87.3	80.9	82.8	87.0	83.0	85.6	87.0	85.9	94.5	59.1	65.3	**93.5**	91.4

Table 2: 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.

LaBSE (Feng et al., 2020), M-MPNET and M-MiniLM (Reimers and Gurevych, 2020), and Jina-Text-v3 (Sturua et al., 2024). Aligned models are denoted as $\mathcal{M}_e \times \mathcal{T}_m$ (e.g., CLIP × LaBSE).

We compare M2M-aligned models against: (i) English-only $\mathcal{M}_e$ (CLIP, Jina-CLIP-v1, K-ALIGN), and (ii) multilingual multimodal baselines (MMMs): mUSEM3L (Aggarwal and Kale, 2020), MCLIP-ST (Multilingual CLIP (Reimers and Gurevych, 2020) from SentenceTransformers[6]), MURAL-Large (Jain et al., 2021), ALIGN-Base (Jia et al., 2021) (via MURAL), XLM-R ViT variants (ViT-L/14, ViT-B/32, ViT-B/16+) & LaBSE ViT-L/14 (Carlsson et al., 2022), Jina-CLIP-v2 (Koukounas et al., 2024b), AltCLIP$_{M9}$ (Chen et al., 2022), SigLIP, SigLIP2[7] (Tschannen et al., 2025), and LLM2CLIP[8] (Huang et al., 2024). Supported languages are listed in Appendix C.

Evaluation. We evaluate on three multilingual image-text 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, 2017; Barrault et al., 2018). Following prior work (Aggarwal and Kale, 2020; Jain et al., 2021; Carlsson et al., 2022), we use Recall@10 with cosine similarity as the ranking score. For XTD, we report Text-to-Image retrieval scores for all languages and the average Recall@10. For XM3600, Multi30K, and Image-to-Text retrieval tasks, we report only the mean Recall@10 across all languages; per-language results are provided in Appendix F.

Results & Analysis. For the XTD Text-to-Image (T2I) task, the M2M-aligned Jina-CLIP-v1 × M-MPNET model (row M4, Table 2) outperforms several multilingual multimodal models trained with multilingual and/or multimodal paired data (rows T1–T3, T5–T7, T11). On English, our aligned models (rows M3 and M4) match the performance of English-trained baselines (rows E1–E3), indicating that alignment does not degrade monolingual performance. Importantly, these gains arise solely from representation alignment, as no

[6] https://www.sbert.net/
[7] HF: google/siglip2-so400m-patch14-384
[8] HF: microsoft/LLM2CLIP-Openai-L-14-336

new multilingual or multimodal supervision is introduced during training. For subsequent comparisons, we adopt Jina-CLIP-v2 as the state-of-the-art reference model, since it achieves the strongest average performance across different languages and datasets.

On XTD, our best M2M-aligned model (row M4) performs 3.1% lower on T2I and 3.8% lower on Image-to-Text (I2T) compared to SOTA. This gap is expected, as models like Jina-CLIP-v2 are explicitly trained on massive multilingual-multimodal data—~400M non-English image-text pairs from CommonPool (Gadre et al., 2023) and 1.2M multilingual synthetic captions. For Multi30K, the performance gap is similar: 3.9% for T2I and 4.3% for I2T. For XM3600, the gap widens to 14.8% (T2I) and 12.6% (I2T), likely due to the larger retrieval space (Multi30K and XTD test sets have 1K instances, while XM3600 has 3,600 images and ~7K captions). Detailed per-language results for XM3600 and Multi30K are provided in Appendix F.

6 Audio-Text Retrieval

Models	AudioCaps		Clotho	
	Avg.	en	Avg.	en
English-only LAION-CLAP Models				
CF: HTSAT-Fused	–	70.3/82.5*	–	**49.9/55.4***
CG: General	–	**83.4**	–	49.3
M2M-aligned Multilingual CLAP models				
CM1: CF × M-MPNET	46.7	62.7	36.7	46.7
CM2: CG × M-MPNET	**54.2**	**77.4**	**36.8**	**47.6**

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

Experimental Setup. We use LAION-CLAP (Wu et al., 2022) as the audio-text multimodal model ($\mathcal{M}_e$) and align it with M-MPNET (T_m). We experiment with two variants of LAION-CLAP: (i) CLAP-HTSAT-fused, trained on AudioCaps (Kim et al., 2019), Clotho (Drossos et al., 2019), and LAION-Audio-630k (Wu et al., 2022); and (ii) CLAP-General, trained on additional speech and music data. For alignment, we use English captions from AudioCaps, Clotho, and WavCaps (Mei et al., 2023). The AudioCaps validation set is used to select the best checkpoint.

Synthetic Evaluation Datasets. Due to the lack of multilingual audio–text evaluation datasets, we extend the AudioCaps and Clotho test sets to 33 additional languages using machine translation. In particular, we translate 4,875 AudioCaps captions (although evaluation was performed on only 4,785 captions because 18 audio files were unavailable during dataset creation) and 5,225 Clotho captions. For 11 Indic languages[9], we use the English-to-Indic translation model from IndicTrans2 (Gala et al., 2023). For the remaining 22 languages[10], we use Aya-23-35B model (Aryabumi et al., 2024).

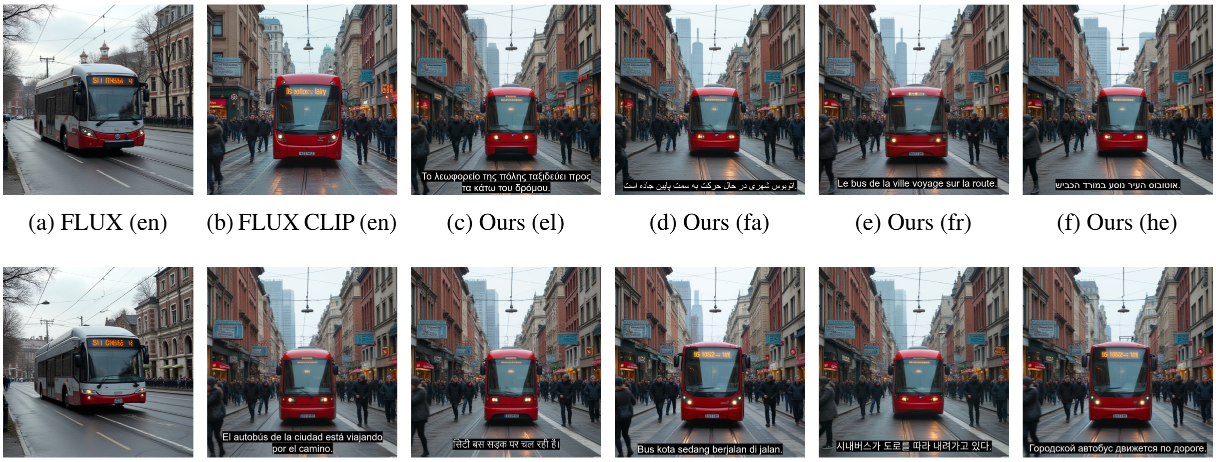
Based on Aya-23-35B's reported results on the FLoRes-200 test set (Costa-jussà et al., 2022) and manual spot checks, we assume that the translations for the 22 languages are of reasonably high quality. The FLoRes-200 test set is also used to identify the optimal prompt for translation. To evaluate translation quality for Indic languages, we back-translate to English using IndicTrans2 (Indic-to-English). Across the 11 Indic languages, AudioCaps achieves a mean spBLEU (Post, 2018) of 48.7 and chrF++ (Popović, 2017) of 63.6, while Clotho achieves 47.4 and 59.6, respectively.

Additional details on dataset licenses and translation quality assessment are provided in Appendix B and G. Due to the lack of comparable multilingual baselines, we report Recall@10 for our method only on these synthetic multilingual test sets. Language-wise Recall@10 scores are provided in Appendix H for both AudioCaps and Clotho.

Results & Analysis. Table 3 shows that our method generalizes effectively to modalities beyond images. On AudioCaps, our approach performs 6% below the SOTA on Text-to-Audio retrieval (T2A), while on Clotho the gap is 2.3% (T2A). To understand this gap, we compute Text-to-Text (T2T) Recall@10 on XM3600 (image-text) and AudioCaps (audio-text), leveraging multiple captions per instance. M-MPNET achieves 62.1% T2T Recall@10 on XM3600, comparable to Jina-CLIP-v1 (63.8%), but only 73.8% on AudioCaps, substantially lower than CLAP-General (80.2%). This suggests that M-MPNET, while effective for image-caption encoding, underperforms for audio-caption encoding.

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

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



(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 (es) (i) Ours (hi) (j) Ours (id) (k) Ours (ko) (l) Ours (ru)

Figure 5: Images generated by FLUX text-to-image model using the prompt "The city bus is traveling down the road" in multiple languages (non-English captions shown on images). Our M2M-aligned model produces similar quality images compared to baseline FLUX (both T5 and CLIP encoders), FLUX-T5 and FLUX-CLIP models.

Qualitative analysis confirms strong semantic alignment. For 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". The ground truth audio ranked 10th, but its captions—"A man talking as metal clacks followed by metal scraping against a metal surface" and "A man is speaking followed by saw blade noises"—closely match the top retrieved results, demonstrating robust semantic retrieval. Additional examples and details are in Appendix H.

7 Cross-lingual Text-to-Image Generation

Our method is task-agnostic and extends naturally to generative tasks such as Text-to-Image generation. Since M2M aligns sentence-level (CLS) representations, we experiment with FLUX.1-dev (FLUX) (Labs, 2024), a 12B parameter Text-to-Image model that conditions on CLS from a CLIP encoder. FLUX is chosen for its public availability, competitive performance (Yang et al., 2024), and dual text encoders: CLIP (CLS conditioning) and T5 (Raffel et al., 2020) (token conditioning).

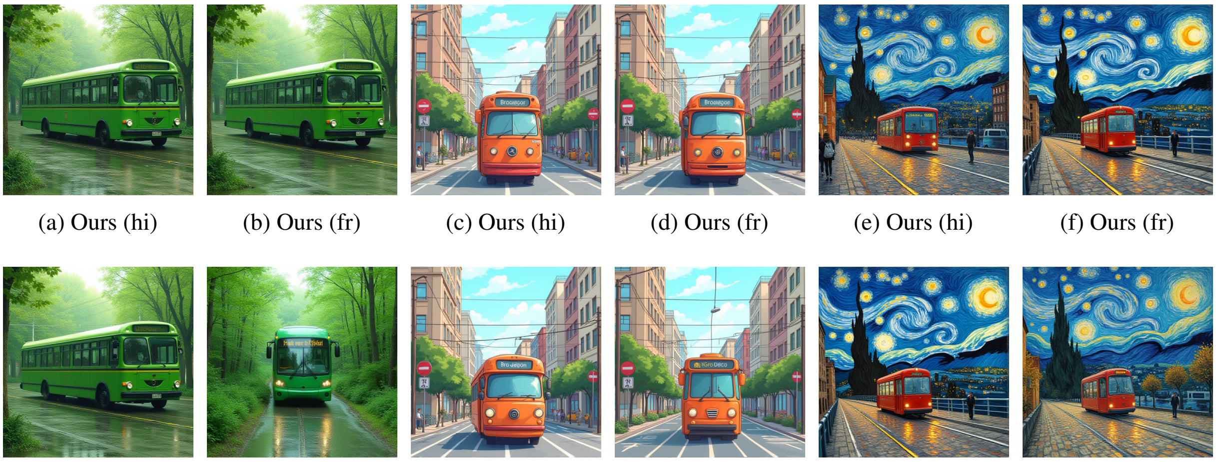
To learn the projection map $\mathcal{F}$, we align the M-MPNET encoder (T_m) with the CLIP encoder from FLUX (T_e).[11] Since FLUX uses both CLIP and T5, we consider four variants: (i) FLUX, which inputs text to both CLIP and T5; (ii) FLUX-CLIP, which inputs text to CLIP and a generic prompt ("A photo of:") to T5[12]; (iii) FLUX-T5, which inputs text to T5 and a generic prompt ("A photo of:") to CLIP; and (iv) FLUX×M-MPNET, which inputs text to the M2M-aligned M-MPNET encoder and a generic prompt to T5.

Training Setup & Evaluation. We follow Section 4, training for 10 epochs without validation, using bfloat16 precision and MSE loss (instead of qq. 6) on unnormalized representations. Unlike retrieval tasks, $\mathcal{L}_{str}$ degrades performance: preserving scale information (exact mapping of representations) rather than structural similarity is more important for generation. Images are generated at 512×512 resolution with guidance scale 3.5, 10 inference steps, and a fixed seed. Following prior work (Ramesh et al., 2021; Rombach et al., 2021; Saharia et al., 2022), we sample 30K captions from MSCOCO2014 (Lin et al., 2014) for English validation set. For multilingual evaluation, we extend captions into 9 languages (fr, el, he, id, ko, fa, ru, es, hi) using IndicTrans2 (hi) and Aya-23-35B (others). Metrics used are FID (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016).

Results & Analysis. FLUX×M-MPNET achieves a strong Inception Score of 31.81 averaged across languages, including 35.9±0.57 on English, sur-

[11] In a qualitative comparison of 100 generations from FLUX×LaBSE vs. FLUX×M-MPNET, the latter consistently produced higher-quality images.

[12] Among tested prompts ("An image of:", "A picture of:", "A photo of:"), the last yielded best results.



(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 6: 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.

passing trained models such as LDM (Rombach et al., 2021) (30.29±0.42), CogView (Ding et al., 2021) (18.2), and LAFITE (Zhou et al., 2022) (26.02). However, FID is poor: 40.9 (FLUX-CLIP) and 43.4 (FLUX×M-MPNET), compared to 23.4 for both FLUX and FLUX-T5. The identical FID for FLUX and FLUX-T5 suggests FLUX relies heavily on T5 token conditioning and can generate high-quality images without CLIP input. Since our setup replaces CLIP with an aligned encoder and uses a generic T5 prompt, generated images are less faithful to the text (e.g., missing objects).

Despite this limitation, qualitative results (Fig. 5) show diverse, semantically relevant images with slightly reduced fidelity. For FLUX-CLIP and FLUX×M-MPNET, we also observe *hallucinated* generations—well-formed but text-misaligned outputs. These are not random noise but coherent, object-rich images, likely caused by weak signal from T5 due to generic prompts. Adding more specific object/style cues to T5 input alleviates this, as illustrated in Fig. 6. Language-wise FID and IS score breakdowns and examples are in Appendix I.

8 Conclusion

We introduce M2M, an efficient method to align multilingual latent spaces with multimodal spaces using only a few linear layers and English text data. Unlike existing approaches requiring large-scale multilingual or multimodal corpora, M2M reduces resource needs while maintaining strong performance across tasks and modalities.

On XTD-T2I retrieval, it achieves 94.9% Recall@10 for English and 89.5% averaged across 11 languages, demonstrating robust zero-shot transfer. Qualitative analyses, including t-SNE visualizations, show projected multilingual embeddings align closely with multimodal representations. Beyond image-text retrieval, M2M generalizes to Audio-Text retrieval and cross-lingual Text-to-Image generation. We release synthetic evaluation datasets: AudioCaps and Clotho extended to 33 languages, and MSCOCO-30K captions extended to 9 languages, providing a unified open benchmark. While promising, further improvements are possible, particularly via token-level alignment. Overall, M2M shows that lightweight, data-efficient strategies can bridge multilingual and multimodal spaces by leveraging implicit alignment between languages and modalities.

Acknowledgements

The authors thank Prof. Preethi Jyothi (Indian Institute of Technology, Bombay) for her continued support and for fostering the research environment that enabled this research. Her guidance and encouragement were instrumental throughout the course of this research. We are grateful to Pavan Kalyan for extensive discussions, careful proofreading, and thoughtful feedback on early drafts of the paper, which greatly improved the clarity, structure, and presentation of this work, and to Isha Pandey for her help in reviewing, refining, and strengthening the rebuttal responses.

Limitations

Need for local alignment. Our method focuses on aligning *global*, sentence-level representations in encoder-based models and demonstrates strong performance in this setting. However, in its current form, it does not provide alignment at the token level and therefore does not directly extend to Multimodal Large Language Models (MLLMs), where effective generation relies on fine-grained representations. Tasks such as text-to-image generation and cross-lingual skill transfer would benefit from token-level alignment signals alongside high-level semantic consistency. Extending the proposed framework to support local alignment is a natural and promising direction for future work.

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*.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Chunyu Wang, Xiyang Dai, Dongdong Chen, et al. 2024. Llm2clip: Powerful language model unlocks richer visual representation. *arXiv preprint arXiv:2411.04997*.

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.

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*.

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, 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.

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[13] 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).

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

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 9). Table 4 shows a list of models and their supported languages.

D Ablations experiments on Image-text Retrieval

Table 5 and 6 shows our method outperforms all other training objectives on Text-to-Image retrieval for XTD dataset. The impact of high λ is significant for both Text-to-Image retrieval and Image-to-Text retrieval (0.5% gain in Avg. Recall@10) shown in Table 6 and Table 5.

E Weight Analysis

To further investigate the behavior of the mapping, we performed a pairwise cosine distance analysis on clusters of semantically identical sentences that differ only in style or phrasing. Distances were computed for (i) Jina-CLIP-v1 embeddings, (ii) multilingual MPNET embeddings, and (iii) our mapped embeddings. All embeddings were ℓ_2-normalized, and only the upper-triangle of the distance matrix was considered to avoid double-counting symmetric pairs. Results show that Jina-CLIP-v1 embeddings exhibit the largest variability (0.028–0.083), reflecting sensitivity to small syntactic or stylistic changes.fig:appendix boxplots Multilingual MPNET embeddings are more compact (0.011–0.043), consistent with their language-agnostic nature, while our mapped embeddings fall in between (0.011–0.046), indicating successful alignment to the Jina-CLIP-v1 space while preserving semantic consistency. These observations support our hypothesis that Jina-CLIP-v1 text encoders capture language-specific features absent in language-agnostic embeddings like M-MPNET, and that the learned linear map effectively projects onto the relevant subspace for retrieval.

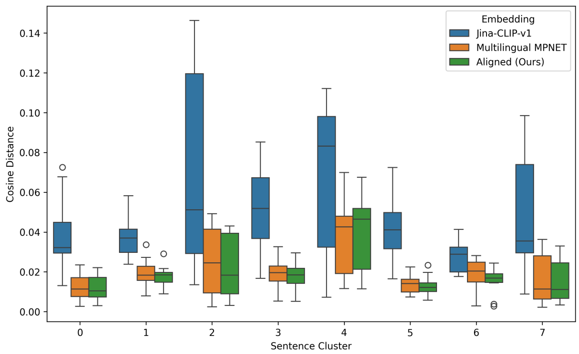


Figure 7: Pairwise cosine distances for clusters of semantically identical sentences with stylistic or syntactic variations. Jina-CLIP-v1 embeddings (blue) show the largest variability, reflecting sensitivity to language-specific phrasing. Multilingual MPNET embeddings (orange) are more compact, consistent with language-agnostic representations, and our mapped embeddings (green) fall in between, indicating successful alignment while preserving semantic consistency. Each boxplot summarizes the distribution of distances within a cluster.

Figure 7 presents boxplots for all clusters, illustrating the distribution of pairwise distances and confirming the trends described above. Below, we list the clusters of nearly identical sentences used for this analysis.

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

Loss	MLP layers	Skip Conn.	Avg.	de	en	es	fr	it	jp	ko	pl	ru	tr	zh
MSE	2	No	89	90.2	94.3	90.8	90.1	90.4	82	85.5	90.8	86	88.9	89.5
MSE	2	Yes	88.9	89.9	94.2	90.3	90	90.7	82	85.4	91	85.9	88.6	89.4
$\lambda_1 * \mathcal{L}_{align} + \beta_1 * \mathcal{L}_{str}$	2	No	**89.5**	90.5	94.7	**91.9**	90.5	91.1	82.4	**85.8**	91.2	**86.8**	**89.2**	89.9
$\lambda_1 * \mathcal{L}_{align} + \beta_1 * \mathcal{L}_{str}$	2	Yes	89.2	90.1	94.3	90.6	90.6	91.1	82.5	85.4	91.1	86.6	88.2	**90.3**
$\lambda_1 * \mathcal{L}_{align} + \beta_1 * \mathcal{L}_{str}$	4	No	89.1	90.3	94.3	90.8	**90.7**	90.8	82.1	85.5	90.7	86.4	88.4	90.2
$\lambda_1 * \mathcal{L}_{align} + \beta_1 * \mathcal{L}_{str}$	1	No	89.3	**90.6**	94.6	90.8	90	90.9	**82.7**	85.7	91	86.6	89	90.1
Similarity Loss	2	No	88.8	89.7	94.2	90.4	89.6	90.5	81.6	84.9	90.8	86	89.2	89.7
L1	2	No	86.7	87.7	94.2	89.2	89.4	89.1	76.9	81.1	88.7	83.7	86.1	87.6
$\lambda_2 * \mathcal{L}_{align} + \beta_1 * \mathcal{L}_{str}$	2	No	89	89.9	**94.8**	90.8	89.8	**91.4**	82.3	84.7	**91.5**	85.7	88.2	89.7

Table 5: 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. $\lambda_1 = 48, \lambda_2 = 1, \beta_1 = 1$.

Loss	MLP layers	Skip Conn.	Avg.	de	en	es	fr	it	jp	ko	pl	ru	tr	zh
MSE	2	No	88.7	89.2	95.3	89.6	89.6	90	81.7	85.4	90.2	85.5	89.4	89.9
MSE	2	Yes	88.6	89.4	95.5	89.8	89.2	89.4	81.9	84.9	90.1	85.5	89.2	89.7
$\lambda_1 * \mathcal{L}_{align} + \beta_1 * \mathcal{L}_{str}$	2	No	**89.4**	**89.6**	95.2	91.2	**89.7**	90.4	**83.1**	**85.9**	90.8	87	90.1	90.3
$\lambda_1 * \mathcal{L}_{align} + \beta_1 * \mathcal{L}_{str}$	2	Yes	**89.4**	89.2	95.5	90.8	89.5	90.5	83	85.5	91	**87.2**	**90.3**	**90.4**
$\lambda_1 * \mathcal{L}_{align} + \beta_1 * \mathcal{L}_{str}$	4	No	89.2	89.3	95.2	90.7	89.5	90.5	83	85.1	90.9	86.8	89.8	**90.4**
$\lambda_1 * \mathcal{L}_{align} + \beta_1 * \mathcal{L}_{str}$	1	No	89.3	89.5	**95.8**	91	89.4	**90.6**	82.4	85.3	**91.1**	86.5	90.1	90.2
Similarity Loss	2	No	88.7	89.2	95.7	90	89.1	89	81.5	85.4	90.5	85.5	89.3	90.1
L1	2	No	83.9	85.6	95	86.2	84.9	87.1	74.9	78.5	84.1	78.7	83.2	85.1
$\lambda_2 * \mathcal{L}_{align} + \beta_1 * \mathcal{L}_{str}$	2	No	88.9	89.2	95.3	**91.3**	**89.7**	**90.6**	82.4	84.3	90.2	86	89.4	90

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 Image-to-Text retrieval task. $\lambda_1 = 48, \lambda_2 = 1, \beta_1 = 1$.

E.1 Sentence Variation Clusters

- **Dog / Animal**
  - A dog running in the park.
  - The dog is running in the park.
  - A dog runs in the park.
  - A dog is running through the park.
  - In the park, a dog runs.
- **Cat**
  - A cat sleeps on the sofa.
  - The cat is sleeping on the sofa.
  - A sleeping cat is on the sofa.
  - On the sofa, a cat is sleeping.
  - A cat is lying asleep on the sofa.
- **Human Actions: Cycling**
  - A person is riding a bicycle on the street.
  - A person is riding a bike along the street.
  - A cyclist is riding on the street.
  - The person rides a bicycle on the street.
  - On the street, a person is riding a bicycle.
- **Human Actions: Drawing**
  - A child is drawing on a piece of paper.
  - The child is drawing on a sheet of paper.
  - On paper, a child is drawing.
  - A child is drawing on paper.
  - The child draws on a piece of paper.
- **Nature / Scenery: Sunset**
  - A sunset over the mountains.
  - The sun is setting over the mountains.
  - The sun is setting near the mountains.
  - The mountains during sunset.
  - The setting sun is over the mountains.
- **Nature / Scenery: River**
  - A river flows through the forest.
  - The river is flowing through the forest.
  - Through the forest flows a river.
  - The forest has a river flowing through it.
  - In the forest, a river flows.
- **Objects / Still Life: Car**
  - A red car is parked on the street.
  - On the street, a red car is parked.
  - A car is parked on the street, and the car is red.

- A car colored red is parked on the street.
- A red-colored car is parked on the street.

- **Objects / Still Life: Coffee**
  - A cup of coffee is on the table.
  - On the table is a cup of coffee.
  - The cup of coffee is on the table.
  - There is a coffee cup on the table.
  - A mug of coffee sits on the table.

F Image-Text Retrieval: Additional Results

F.1 Language-wise Recall on XM3600 & Multi30K

Tables 7, 8 show language-wise performance of our M2M-aligned models on XM3600 and Multi30K datasets respectively. Interestingly, CLIP × M-MPNET outperforms Jina-CLIP-v1 × M-MPNET by 2.4% I2T and 0.2% T2I on Multi30K dataset.

F.2 Results on model-supported languages

Similar to our results for Image-Text retrieval in Table 2, in Table 9, 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 4.

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 10 to show that the performance is stable across varying random seeds.

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[14], 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 13, and for backtranslated Indic translations are shown in Table 14. Zero-shot prompt and 3-shot prompt templates are listed in Table 11 and 12.

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 15 and Clotho in Table 16.

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 17.

I Cross-lingual Text-to-Image Generation.

Both Inception score and FID scores are computed using torch-fidelity (Obukhov et al., 2020) package [15]. Language-wise FID scores shown in Table 19 and inception scores are shown in Table 18.

[14] ar, zho-Hant, fr, de, he, hi, it, jp, ko, pl, ru, es, tr, vi
[15] https://github.com/toshas/torch-fidelity

Retrieval Type	Avg	ar	bn	cs	da	de	el	en	es	fa	fi	fil
T2I	66.3	68.3	31.3	73	79.3	80.9	67.3	79.8	75.7	76	77.3	9.8
I2T	73.1	77.4	36.6	80.2	86.6	88	76.1	84.9	82.4	82	84.8	17.9

Retrieval Type	fr	he	hi	hr	hu	id	it	ja	ko	mi	nl	no
T2I	81.4	74.7	60	79.4	77.7	85.5	79.4	78.7	72.4	0.7	75	78.9
I2T	88.3	82.6	70.7	87.3	83.5	90	85.8	85.6	81.2	1.1	80	86.9

Retrieval Type	pl	pt	quz	ro	ru	sv	sw	te	th	tr	uk	vi	zh
T2I	76.2	77.2	2.7	80	82.2	78	4.5	28.8	78.9	74.9	76.8	81.9	80.9
I2T	83.6	84.1	6.1	87.6	88.6	85.3	9.1	39.1	86.1	81.4	84	89.2	87.4

Table 7: 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	89.9	87.9	**89.5**	91.8	90.3	89.7	86.8	89.7	91.7	90.4
CLIP × M-MPNET	**90.1**	**88.1**	88.3	**93.2**	**90.8**	**92.1**	**90**	**90.9**	**95.1**	**92.3**

Table 8: 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	64.9	67.5	79	75.7
M2: Jina-CLIP-v1 × M-MiniLM	68.5	75.7	88	85.9
M3: Jina-CLIP-v1 × JinaTextV3	75.4	80.1	87.8	87.5
M4: Jina-CLIP-v1 × M-MPNET	**76.8**	**84.0**	89.9	89.7
M5: CLIP × M-MPNET	65.4	77.6	90.1	**92.1**
M6: K-ALIGN × M-MPNET	68.6	78.3	**91.0**	90.2

Table 9: 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.

Retr. Type	Avg.	de	en	es	fr	it	jp	ko	pl	ru	tr	zh
T2I	89.4±0.1	90.6±0.1	94.6±0.1	91.5±0.6	90.4±0.1	91.1±0.0	82.7±0.4	85.5±0.4	91.1±0.1	86.8±0.1	89.4±0.3	89.8±0.1
I2T	89.4±0.1	89.2±0.6	95.4±0.2	91.0±0.3	89.6±0.1	90.4±0.0	83.1±0.0	85.7±0.4	91.0±0.3	87.0±0.0	90.1±0.0	90.4±0.1

Table 10: 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-Text retrieval task on XTD dataset.

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 11: 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 12: 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.

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 8, Figure 9 & Figure 10.

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 13: 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 14: 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	48.5	46.2	23.3	57.6	58.5	55.6	77.4	60.8	38.4	48.7	52.9	59	58.2
A2T	61	61.2	35.3	66.2	68.7	69.5	81.6	70.2	53	61.3	63.4	68.7	67.4

Retrieval Type	ja	kn	ko	ml	mr	nl	ne	pa	fa	pl	pt	ro	ru
T2A	50.2	24	48.5	26.4	46.3	60.8	33.8	22.7	54.6	52.5	61.3	58.2	49.8
A2T	66.4	39.1	63.3	42.6	61.9	70.6	47.1	35.4	69.7	66.4	69.8	68	64.8

Retrieval Type	es	ta	te	tr	uk	ur	vi	zh (hans)	zh (hant)
T2A	59.6	29.3	25.7	56.7	46.8	44.7	56.6	53.3	51.1
A2T	67.9	42.9	38.5	64.8	62.8	58	68.3	70.6	67.9

Table 15: 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 × M-MPNET model.

Retrieval Type	Avg	ar	bn	cs	de	el	en	fr	gu	he	hi	id	it
T2A	33.3	34.2	19	37.5	37.8	36	47.6	40.1	25.9	33.7	36.6	39.6	38.1
A2T	39.5	42	24.4	41.8	42.5	41.3	50.7	44.8	34.2	39	42.9	44.7	44.7

Retrieval Type	ja	kn	ko	ml	mr	nl	ne	pa	fa	pl	pt	ro	ru
T2A	38.8	17.2	35	20.9	29.2	38.6	23.8	18	37	36.6	39.2	38.3	35.5
A2T	45.6	25.5	43	27.7	36.5	45.1	31.3	25.1	43.3	43.2	43.9	43.6	43.8

Retrieval Type	es	ta	te	tr	uk	ur	vi	zh (hans)	zh (hant)
T2A	39.8	20.8	17.5	36.3	33.4	31.2	39	39.9	39.6
A2T	45.5	27.9	24.5	43.2	41.1	38.2	44.1	45.4	43.9

Table 16: 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 × M-MPNET model.

	Captions of Retrieved Audios			
Query Text	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

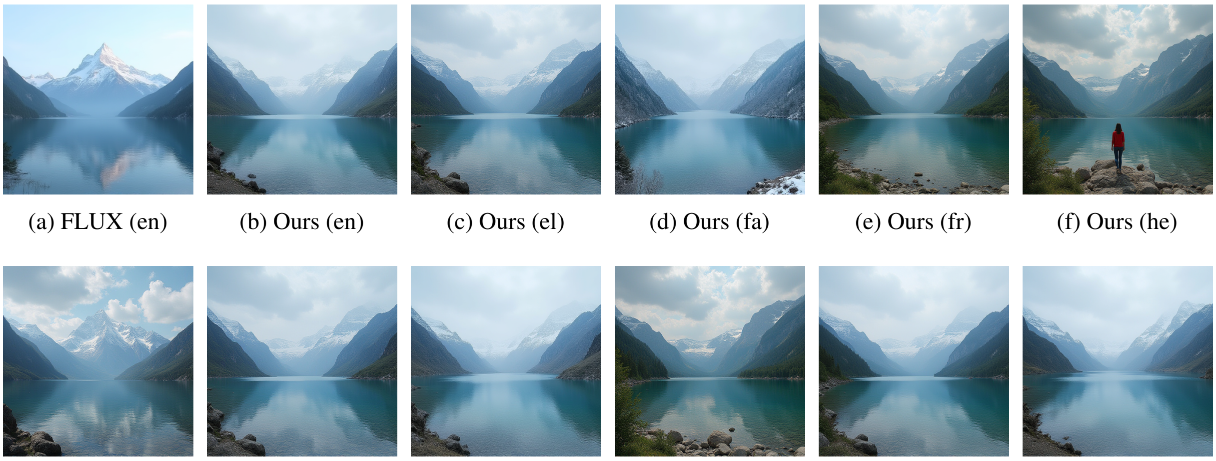
Table 17: Captions of Audio retrieved for a Query text (Text-to-Audio retrieval task) using M2M-aligned CLAP-General × M-MPNET

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

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 19: 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 8: 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 9: 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.

Figure 10: 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).