HanCLIP: Korean Vision-Language model via Connecting Contrastive Representations

Korea University COSE461 Final Project

Seunggi Moon
Department of Computer Science
Team 6
2020320071

Yujin Sung
Department of Computer Science
Team 6
2022320014

Donghyun Shin
School of Mechanical Engineering
Team 6
2020170616

Dahyun Chung
Division of Life Science
Team 6
2019140004

Abstract

Contrastive Language-Image Pre-training (CLIP) have demonstrated impressive generalization across a wide range of downstream tasks by learning from massive paired English-image data. However, while English text and images are well aligned, achieving similar alignment in other languages remains challenging due to the scarcity of high-quality text-image pairs in many languages, including Korean. Directly training multilingual vision-language models typically requires extensive annotation efforts or costly fine-tuning of large-scale encoders. To address this challenge, we propose a training-efficient, paired-data-free framework that enables Korean–image alignment without modifying or fine-tuning any pretrained encoder. Our method leverages a multilingual text encoder and the pretrained CLIP model to construct a Korean–image contrastive representation space, using English as a shared semantic bridge. Specifically, we utilize English text embeddings as anchors that connect Korean text and image representations. Since both Korean–English translation pairs and English–image pairs are widely available or already embedded, we can implicitly associate Korean text and images by aligning them through this common English modality. We train a lightweight projection module to align Korean and image embeddings in a shared space using a contrastive loss defined over the English bridge. This approach preserves the semantic structure of the original CLIP space while avoiding the need for paired Korean–image data and keeping the vision encoder entirely frozen, resulting in efficient and scalable cross-lingual vision-language alignment.

1 Introduction

Contrastive vision-language models, particularly those based on architectures like CLIP [27], have set new benchmarks across a wide range of visual understanding tasks, owing to their strong generalization capabilities. This success is largely driven by large-scale paired image-text datasets—predominantly in English—which facilitate effective alignment between visual and textual modalities. However, such resources are scarce in other languages, including Korean, creating a significant data imbalance that limits the global applicability and fairness of vision-language technologies. Bridging this language gap is increasingly important amid rising demand for multilingual

Preprint. Under review.

AI services such as cross-lingual information retrieval [23], content recommendation [25], digital media analysis [2], and personalized user experiences [24].

A common approach is to train multilingual vision-language models directly on paired image–text data in the target language [13, 14, 36]. However, this strategy is hindered by the high cost of constructing large-scale datasets and the computational burden of fine-tuning large models. Moreover, existing multilingual extensions like Multilingual-CLIP [5] often underperform due to weak alignment between language-specific and visual representations and the limited availability of high-quality parallel corpora.

Although Korean is widely considered high-resource in NLP [17, 16, 26], it lacks large-scale, high-quality multimodal corpora necessary for training robust vision-language models. For instance, CLIP was trained on over 400M English image-text pairs, while Korean datasets like AIHub's MSCOCO-Kor [1] contain only about 120K samples, many of which are machine-translated or domain-limited. This stark data disparity undermines transferability and prevents existing CLIP-based models from generalizing well to Korean domain. A commonly adopted workaround is to translate Korean text into English and query pretrained CLIP models. However, this two-stage pipeline introduces latency, semantic distortion, and modality misalignment—especially when visual context is lost during translation or when ambiguous terms are inadequately resolved.

Recent work, such as C-MCR [33], has shown that semantic anchors like text can bridge disjoint modalities (e.g., image–audio). Inspired by this, we propose an efficient, scalable framework to align Korean text and image representations—**without requiring any paired Korean–image data with the pretrained encoder frozen**. Our method, *HanCLIP*, leverages the abundance of Korean–English parallel text and the strong alignment between English text and images from pretrained CLIP. Instead of requiring direct Korean–image supervision, we use English as a shared semantic bridge to implicitly connect Korean and image embeddings.

The proposed framework comprises three components: (1) a frozen multilingual text encoder that maps Korean and English into a shared space, (2) a frozen CLIP model that provides English–image alignment, and (3) lightweight projection heads trained to align Korean and image representations via contrastive learning over shared English anchors. Notably, our projection module consists of only **2.1M trainable parameters**—over **99.5% fewer** than KoCLIP [18] (425M) and **99.6% fewer** than Multilingual-CLIP (560M)—making it significantly more compact than full CLIP models and enabling efficient training and inference on modest hardware. Furthermore, compared to the commonly adopted CLIP+Translator [9] pipeline, which requires separate translation and image encoding stages, *HanCLIP* achieves more than **90% faster inference latency**, making it well-suited for real-time and resource-constrained applications.

Our contributions are as follows:

- We propose *HanCLIP*, a novel cross-lingual vision-language alignment framework that bridges Korean and image modalities via English as a semantic pivot, without requiring any direct Korean–image supervision.
- Our method preserves the pretrained CLIP and multilingual text encoder, and trains only a lightweight projection module (2.1M parameters), achieving efficient alignment with minimal computational overhead—over 99% smaller than prior baselines.
- Despite using no paired Korean–image data, *HanCLIP* demonstrates strong zero-shot performance on multiple Korean–image retrieval and classification benchmarks, showing effective and transferable representation across modalities and languages.
- *HanCLIP* reduces inference latency by more than 10 times compared to CLIP+Translator pipelines, enabling fast deployment in real-world multilingual systems.

2 Related Work

2.1 Vision-Language Pretrained Models

Vision-Language Pretrained Models (VL-PTMs) can be broadly categorized based on how they model interactions between image and text modalities. Fusion encoder models jointly process multimodal inputs, either through a single-stream architecture [11, 21, 34] or dual-stream designs with cross-attention modules [12, 20, 35]. In contrast, dual encoder models such as CLIP [27], ALIGN [15],

and DeCLIP [29] encode image and text inputs independently via modality-specific encoders and map them into a shared embedding space. This design allows efficient similarity computation and scalability for large-scale retrieval tasks. Hybrid models [3, 28] have also emerged, combining the strengths of both fusion and dual encoder architectures. However, the success of VL-PTMs fundamentally depends on the availability of large-scale paired image–text data, which remains predominantly English-centric. This reliance poses a major challenge when adapting to languages with limited multimodal resources, such as Korean.

2.2 Cross-Lingual and Multilingual Extensions

To expand cross-lingual applicability, several efforts [36, 18] have adapted CLIP-like models to support non-English languages. Approaches such as Multilingual-CLIP [5] incorporate multilingual supervision, often requiring substantial language-specific paired datasets and fine-tuning of encoder components. MURAL [13] extends this line by aligning multiple languages with English via shared image anchors to support multilingual retrieval. Despite these advances, most methods still depend heavily on large-scale annotated data for each target language, which restricts their use in truly low-resource scenarios. These limitations motivate the development of alternative approaches that can exploit high-resource pivot languages, such as English, to enable cross-lingual vision-language alignment without direct supervision in the target language.

2.3 Indirect Cross-Modal Alignment

Recent works on cross-modal representation learning have proposed lightweight and indirect alignment frameworks that minimize parameter overhead while avoiding direct supervision between modality pairs. For instance, C-MCR [33] introduces a contrastive objective with shared text anchors to align modality pairs such as image and audio. Similarly, mCLIP [6] distills knowledge from CLIP into a multilingual encoder using lightweight projection heads, thereby reducing the need for full model fine-tuning.

3 Approach

3.1 Overall Approach

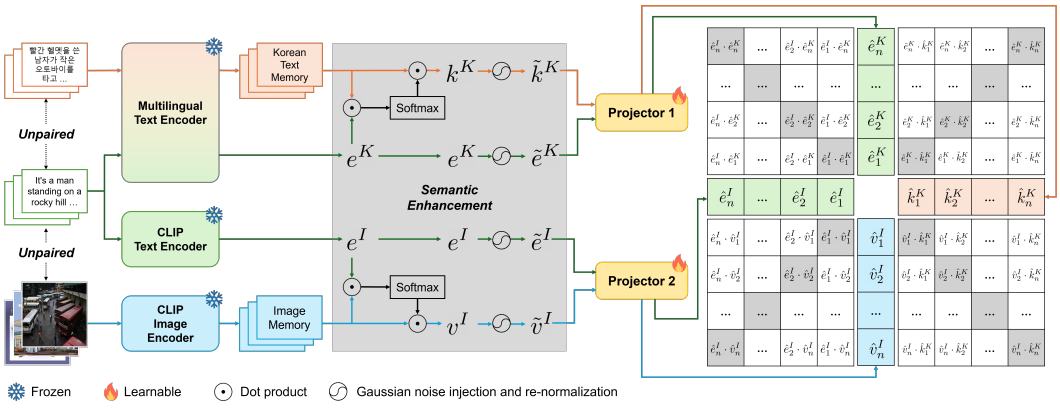

Figure 1: **Overall Architecture.**

HanCLIP aligns Korean text and images without paired data or encoder fine-tuning, using English as a semantic bridge. It comprises frozen encoders (CLIP and multilingual) and two lightweight projection heads. Given an English query, pseudo-aligned Korean and image embeddings are retrieved from memory banks, perturbed with noise, and projected into a shared space. Training optimizes inter- and intra-alignment losses to build semantic consistency across modalities.

3.2 Semantic Enhancement

Cross-modal and Cross-lingual Semantic Enhancement Given an English query e_i, we obtain two embeddings: e_i^I from the CLIP text encoder and e_i^K from a multilingual text encoder. We construct two memory banks: image memory $V = \{v_1, \ldots, v_N\}$ encoded by the CLIP image encoder, and Korean text memory $K = \{k_1, \ldots, k_M\}$ encoded by the same multilingual text encoder.

The image representation aligned with the English query is retrieved as:

$$v_i^I = \sum_{k=1}^{N} \frac{\exp(\text{sim}(e_i^I, v_k)/\tau_1)}{\sum_{j=1}^{N} \exp(\text{sim}(e_i^I, v_j)/\tau_1)} \cdot v_k \tag{1}$$

Likewise, the Korean text embedding aligned with the same query is:

$$k_i^K = \sum_{k=1}^{M} \frac{\exp(\text{sim}(e_i^K, k_k)/\tau_1)}{\sum_{j=1}^{M} \exp(\text{sim}(e_i^K, k_j)/\tau_1)} \cdot k_k \tag{2}$$

These soft retrievals yield pseudo-aligned features v_i^I and k_i^K, bridging modalities through the shared semantics of the English query without requiring direct Korean–image supervision.

Perturbed Embedding Semantic Enhancement Inspired by previous work [33], to enhance robustness against encoding biases, we add Gaussian noise to the four embeddings e^I, e^K, v^I, k^K, and normalize them onto the unit hypersphere:

$$\tilde{e}_i^I = \text{Normalize}(e_i^I + \theta_1), \quad \tilde{e}_i^K = \text{Normalize}(e_i^K + \theta_2) \tag{3}$$

$$\tilde{v}_i^I = \text{Normalize}(v_i^I + \theta_3), \quad \tilde{k}_i^K = \text{Normalize}(k_i^K + \theta_4) \tag{4}$$

where $\theta_1, \theta_2, \theta_3, \theta_4 \sim \mathcal{N}(0, \sigma^2 I)$. This augmentation promotes local semantic smoothness and trains the model to align perturbed embedding distributions, improving generalization and stability across modalities.

3.3 Inter-alignment

For each English query i, we first obtain four perturbed embeddings: $\tilde{e}_i^I$ and $\tilde{e}_i^K$ denote the representations of the same English query obtained from the CLIP text encoder and the multilingual text encoder, respectively; $\tilde{v}_i^I$ and $\tilde{k}_i^K$ are the image and Korean text features retrieved from the memory banks using these text embeddings. Each of these four embeddings is projected into a shared embedding space via the corresponding projection heads as follows:

$$\hat{e}_i^I = f_1(\tilde{e}_i^I), \quad \hat{e}_i^K = f_2(\tilde{e}_i^K)$$
$$\hat{v}_i^I = f_1(\tilde{v}_i^I), \quad \hat{k}_i^K = f_2(\tilde{k}_i^K)$$

The pair $(\hat{e}_i^I, \hat{e}_i^K)$, representing the same English query embedded by two different encoders, provides a strong supervision signal to align textual semantics across encoders. On the other hand, the pair $(\hat{v}_i^I, \hat{k}_i^K)$, which associates retrieved image and Korean text embeddings, serves as a pseudo-aligned cross-modal pair.

To align these representations, we apply symmetric contrastive losses to both pairs. The text-text alignment loss is defined as:

$$\mathcal{L}_{\text{text}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(\text{sim}(\hat{e}_i^I, \hat{e}_i^K)/\tau_2)}{\sum_{j=1}^{B} \exp(\text{sim}(\hat{e}_i^I, \hat{e}_j^K)/\tau_2)} + \log \frac{\exp(\text{sim}(\hat{e}_i^K, \hat{e}_i^I)/\tau_2)}{\sum_{j=1}^{B} \exp(\text{sim}(\hat{e}_i^K, \hat{e}_j^I)/\tau_2)} \right] \tag{5}$$

Similarly, the pseudo image–Korean pair alignment loss is given by:

$$\mathcal{L}_{\text{pseudo}} = -\frac{1}{2B} \sum_{i=1}^{B} \left[\log \frac{\exp(\text{sim}(\hat{v}_i^I, \hat{k}_i^K)/\tau_3)}{\sum_{j=1}^{B} \exp(\text{sim}(\hat{v}_i^I, \hat{k}_j^K)/\tau_3)} + \log \frac{\exp(\text{sim}(\hat{k}_i^K, \hat{v}_i^I)/\tau_3)}{\sum_{j=1}^{B} \exp(\text{sim}(\hat{k}_i^K, \hat{v}_j^I)/\tau_3)} \right] \tag{6}$$

The final inter-alignment loss is defined as the sum of these two:

$$\mathcal{L}_{\text{inter}} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{pseudo}} \tag{7}$$

The alignment between $\hat{e}_i^I$ and $\hat{e}_i^K$, derived from the same English query, acts as a strong supervisory signal. This supervision helps guide the weaker, pseudo-aligned pair $(\hat{v}_i^I, \hat{k}_i^K)$, facilitating robust cross-modal and cross-lingual representation learning.

3.4 Intra-alignment

Due to modality gaps, semantically aligned embeddings like $\hat{e}_i^I$ and $\hat{v}_i^I$, or $\hat{e}_i^K$ and $\hat{k}_i^K$, often reside in different regions. Inspired by prior work[22], we mitigate this by removing the repulsive term from contrastive loss and retaining only the attractive component.

Standard contrastive loss decomposes as:

$$-\log \frac{\exp(\text{sim}(x_i, z_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(x_i, z_j)/\tau)} = -\frac{1}{\tau}\text{sim}(x_i, z_i) + \log \sum_{j=1}^{N} \exp\left(\frac{\text{sim}(x_i, z_j)}{\tau}\right) \tag{8}$$

We retain only the attractive term and define the intra-alignment loss:

$$\mathcal{L}_{\text{intra}} = \frac{1}{2B} \sum_{i=1}^{B} \left(\|\hat{e}_i^I - \hat{v}_i^I\|_2^2 + \left\|\hat{e}_i^K - \hat{k}_i^K\right\|_2^2 \right) \tag{9}$$

As all embeddings are ℓ_2-normalized, the squared distance becomes:

$$\|x - y\|^2 = 2(1 - x^\top y) \tag{10}$$

This encourages angular proximity and closes modality gaps within both CLIP and multilingual spaces. The intra-alignment complements inter-alignment, enabling robust Korean–image representation alignment via English intermediaries.

3.5 Training and Inference

During training, all pretrained encoders—including the CLIP image/text encoders and the multilingual text encoder—are frozen. Only the projection networks $f_1(\cdot)$ and $f_2(\cdot)$, which map embeddings into a shared space, are optimized. To improve efficiency, we precompute the English embeddings e_i^I and e_i^K, as well as the retrieved embeddings v_i^I and k_i^I, since the retrieval is training-free and based on similarity. This avoids repeated forward passes through the encoders during optimization. The projection heads are trained using both **inter-** and **intra-alignment** losses. Inter-alignment ensures cross-modal and cross-lingual consistency, while intra-alignment reduces the modality gap. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{inter}} + \lambda \mathcal{L}_{\text{intra}}, \tag{11}$$

where λ balances the two terms. This enables *HanCLIP* to learn a unified space where Korean text and images are indirectly aligned via English semantics. During inference, the CLIP image embedding and multilingual Korean text embedding are projected via their respective heads. Cosine similarity in the shared space determines semantic relevance, enabling flexible Korean–image retrieval without direct Korean–image supervision during training.

4 Experiments

4.1 Datasets

To construct the English text memory, we sample a total of 1M English captions from COYO captions[4]. The captions are unpaired with Korean text or images during training and are used to enhance the semantic diversity and robustness of the shared embedding space. For the Korean text memory, we sample 0.6M Korean captions from MSCOCO-Korean dataset[1]. The dataset contain human-written or machine-translated captions describing natural images and are used to construct the Korean side of the multilingual text embedding space. The Korean captions are unpaired with English text or images. The image memory is built from 1.5M images randomly sampled unlabeled images from ImageNet-1K[8]. These images are not paired with either English or Korean captions during training and serve as the visual modality for Korean-visual downstream tasks.

Method	I-K Pairs	Trainable Params	MSCOCO-Kor[1]						KoCC3M[30]					
			I→K			K→I			I→K			K→I		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Multilingual-CLIP[5]	✓	560M	0.5781	0.8437	0.9218	0.6265	0.9000	0.9562	0.0156	0.0312	0.0313	0.0062	0.0359	0.0843
KoCLIP[18]	✓	425M	0.6015	0.9062	0.9609	0.6312	0.8921	0.9187	0.0234	0.0546	0.0547	0.0156	0.0374	0.0671
HanCLIP (Ours)	×	**2.1M**	0.6563	0.9297	0.9687	0.5859	0.8859	0.9453	0.0178	0.0625	0.1016	0.0047	0.0375	0.0781

Table 1: **Bidirectional Korean-Image retrieval results on MSCOCO-Kor[1] and KoCC3M[30].** "I-K Pairs" indicates whether the model was trained with paired Image–Korean data. "Trainable Params" refers to the number of trainable parameters. I→K denotes Image-to-Korean retrieval and K→I denotes Korean-to-Image retrieval. Best and second-best scores are highlighted.

4.2 Implementation details

We adopt a frozen CLIP ViT-B/32 model as the vision encoder and use MiniLM-L12 as the multilingual text encoder, both of which remain frozen during training. Among the candidate encoders (MiniLM-L12 and E5-base), we select MiniLM-L12 based on the ablation results in Table 3, where it shows superior performance on Korean-visual downstream tasks. The two projectors $f_1(\cdot)$ and $f_2(\cdot)$ are implemented as simple multi-layer perceptrons consisting of two linear layers with BatchNorm and ReLU activation, as described in Table 5. The temperature parameters τ_1, τ_2 and τ_3 used in Equations (1), (2), (5) and (6) are all set to $1/1000$. For noise injection in Equation (3), (4) the Gaussian noise variance σ^2 is set to 0.004. The balancing coefficient λ in the final loss function Equation (11) is set to 0.1. We train the projectors for 36 epochs using a batch size of 4. The AdamW optimizer is used with an initial learning rate of $1e-3$, and the learning rate is scheduled with cosine decay. All experiments are conducted with frozen encoders, and only the projection layers are updated during training.

4.3 Evaluation

Due to the limited availability of open-source CLIP variants tailored for Korean-language understanding, we compare *HanCLIP* against two leading publicly available baselines: Multilingual-CLIP [5] and KoCLIP [18]. Multilingual-CLIP is trained on translated datasets including 118K MS-COCO, 3.3M Google Conceptual Captions (GCC), and 23.4K VizWiz, totaling about 3.5M Korean–image pairs. KoCLIP is trained on roughly 120K Korean–image pairs from the MSCOCO-Kor dataset. In contrast, *HanCLIP* is trained without any Korean–image supervision and requires only 2.1M trainable parameters—over 200 times fewer than KoCLIP and nearly 280 times fewer than Multilingual-CLIP. *HanCLIP* is evaluated on image–Korean retrieval, Korean–image retrieval, and zero-shot classification.

For image-to-Korean retrieval, a query image is embedded using the CLIP image encoder and projected via $\hat{v}_q = f_1(v_q)$. Korean candidate sentences $\{\hat{k}_1, \ldots, \hat{k}_M\}$ are encoded with a multilingual text encoder and projected via $f_2(\cdot)$. Cosine similarities $\text{sim}(\hat{v}_q, \hat{k}_j)$ are computed to retrieve the top-matching captions. We evaluate on MSCOCO-Korean [1] and KoCC3M [30] using Recall@1, 5, and 10.

In the Korean-to-image setting, a Korean query is projected via $\hat{k}_q = f_2(k_q)$ and compared against precomputed image embeddings $\{\hat{v}_1, \ldots, \hat{v}_N\}$. Images are ranked by cosine similarity $\text{sim}(\hat{k}_q, \hat{v}_j)$ and evaluated on the same benchmarks.

We also assess zero-shot classification on four Korean-translated benchmarks: CIFAR-10 [19], STL-10 [7], CIFAR-100 [19], and Caltech101 [10]. Each image is encoded via the CLIP encoder and projected using $f_1(\cdot)$; class names are encoded and projected via $f_2(\cdot)$. Classification is performed via cosine similarity, and predictions are based on the most similar class. We report average F1 scores to account for class imbalance. As these datasets are unseen during training, the results demonstrate the strong zero-shot generalization of *HanCLIP*.

4.4 Results

Table 1 presents results on MSCOCO-Kor [1] and KoCC3M [30] benchmarks. On image-to-Korean retrieval, *HanCLIP* achieves the highest scores across all recall levels on MSCOCO-Kor, with Recall@1 of 0.6563, surpassing KoCLIP (0.6015) and Multilingual-CLIP (0.5781) despite using no

Method	I-K Pairs	Trainable Params	CIFAR-10[19]	STL-10[7]	CIFAR-100[19]	Caltech101[10]
Multilingual-CLIP[5]	✓	560M	0.5922	0.8294	0.2712	0.0933
KoCLIP[18]	✓	425M	0.4348	0.6330	0.0779	0.0522
HanCLIP (Ours)	×	**2.1M**	0.7297	0.8137	0.2412	0.0823

Table 2: **Zero-shot image classification results on Korean-translated datasets.** "I-K Pairs" indicates whether the model was trained with paired Image–Korean data. "Trainable Params" refers to the number of trainable parameters. We report average F1 scores on CIFAR-10[19], STL-10[7], CIFAR-100[19], and Caltech101[10], where the class labels are translated into Korean. Best and second-best scores are highlighted.

paired data and only 2.1M trainable parameters. On KoCC3M, a fully unseen benchmark, it achieves Recall@10 of 0.1016, nearly doubling KoCLIP and outperforming Multilingual-CLIP by over 3 times, demonstrating strong generalization to new data.

In the Korean-to-image retrieval task, *HanCLIP* performs comparably to baselines on MSCOCO-Kor, with Recall@10 of 0.9453. While slightly lower on Recall@1 and Recall@5, the performance gap remains small. On KoCC3M, *HanCLIP* achieves the highest Recall@5 (0.0375) and strong Recall@10 (0.0781), showing consistent zero-shot robustness. Qualitative retrieval examples are provided in Appendix Figure 4.

Table 2 shows zero-shot classification performance on CIFAR-10[19], STL-10[7], CIFAR-100[19], and Caltech101[10], using Korean-translated labels. *HanCLIP* achieves the best average F1 score on CIFAR-10 (0.7297), outperforms KoCLIP and Multilingual-CLIP on most benchmarks, and remains competitive even on fine-grained datasets. These results confirm that *HanCLIP*'s compact, supervision-efficient alignment generalizes beyond retrieval.

5 Analysis

5.1 Visualization of Cosine Similarity

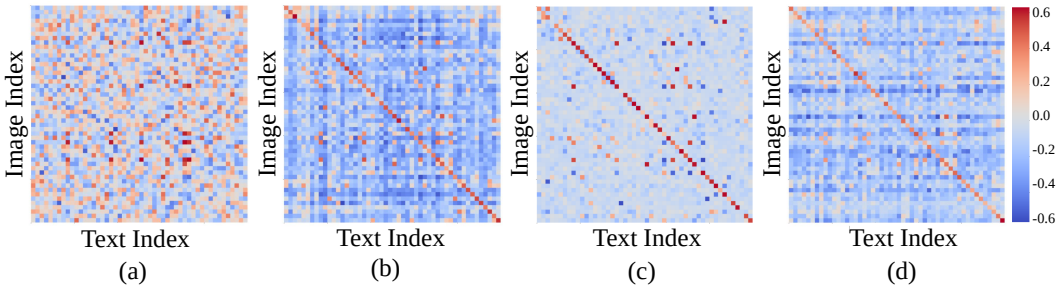

Figure 2: **Cosine similarity visualization for embeddings of Korean and image queries.** We visualize the cosine similarity matrices under four different settings: (a) before projection, (b) using our proposed *HanCLIP* model, (c) using the publicly released KoCLIP model[18], and (d) using Multilingual-CLIP model[5]. All text samples are in Korean.

To evaluate Korean–image alignment, we visualize cosine similarity matrices between 50 image–text pairs from MSCOCO-Kor[1] under four settings. In Figure 2(a), raw image and Korean text embeddings show no clear structure, indicating poor alignment. In contrast, *HanCLIP* (Figure 2(b)) exhibits a sharp diagonal with low off-diagonal values, reflecting strong semantic consistency and accurate alignment. KoCLIP (Figure 2(c)) also shows a diagonal pattern, but with more diffusion and higher off-diagonal similarity, suggesting weaker discrimination between semantically similar but incorrect pairs. Multilingual-CLIP (Figure 2(d)) reveals a fragmented diagonal and residual off-diagonal noise—better than raw features, but less precise than *HanCLIP*. This may stem from limited Korean-specific supervision or its language-agnostic design. Overall, Figure 2 highlights the superiority of *HanCLIP* in aligning Korean-language text with visual content. Compared to both unprojected features and baselines such as KoCLIP[18] and Multilingual-CLIP[5], *HanCLIP*

produces the most distinct, structured, and reliable joint representations. This strongly supports its effectiveness for Korean-visual downstream tasks, where fine-grained semantic alignment is crucial.

5.2 Visualization of image embedding

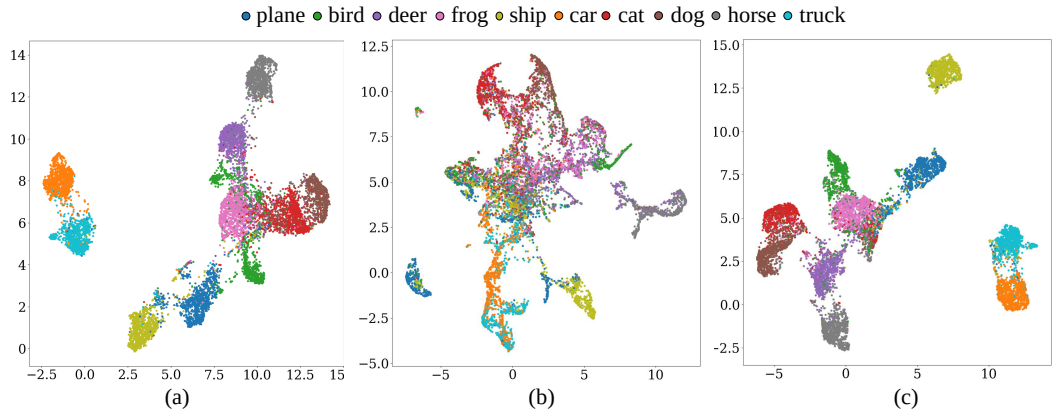

Figure 3: **UMAP visualization of image embedding.** We visualize embeddings extracted from (a) the CLIP [27] image encoder, (b) the KoCLIP [18] image encoder, and (c) the *HanCLIP* model, where a trained projection layer is applied to the CLIP image features. Results are based on CIFAR-10 [19].

To qualitatively assess the effectiveness of visual representations across different models, we visualize image embeddings from CIFAR-10 [19] using UMAP. We exclude Multilingual-CLIP [5], which freezes the CLIP image encoder and adapts only the text encoder, resulting in identical image embeddings to CLIP. Embeddings from CLIP (Figure 3(a)) form well-separated clusters that align closely with semantic categories, demonstrating the strong representational power of the pretrained CLIP encoder. In contrast, the KoCLIP image encoder (Figure 3(b)) shows substantial overlap across all categories, with class boundaries largely collapsed. This indicates that the visual representation lacks clear semantic separability, suggesting significantly reduced discriminability. Figure 3(c) shows that *HanCLIP* preserves clear category separation, even after applying a trained projection layer to the CLIP image features. Clusters remain compact and distinct, particularly for visually similar classes such as *cat*, *dog* and *truck*, *car*. This confirms that *HanCLIP* successfully inherits CLIP's visual representations, and that the added lightweight projection layer preserves, not degrades, the integrity of the image embedding space for downstream tasks.

6 Conclusion

6.1 Limitations

While *HanCLIP* shows strong performance in zero-shot retrieval and classification, several limitations remain. First, it relies on pre-trained multilingual encoders whose quality is bounded by the coverage and bias of their training corpora. As a result, culturally specific expressions, informal speech, or underrepresented Korean domains may not be well captured. Second, although our approach enables alignment without paired Korean–image data, the training datasets mainly consist of globally generic content. This may limit grounding of fine-grained Korean semantics. Lastly, our evaluation is limited to retrieval and classification; extending *HanCLIP* to generation and reasoning tasks remains for future work.

6.2 Future Work

A promising direction is to generalize *HanCLIP* to other non-English languages by leveraging stronger multilingual encoders and advances in cross-lingual transfer. We also aim to adapt the model to generation tasks such as captioning and VQA, enabling more complex cross-modal understanding. Finally, incorporating culturally rich Korean data—both visual and textual—can help *HanCLIP* become not only linguistically aligned, but also culturally grounded.

References

[1] AIHub. Korean image captioning dataset. https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&dataSetSn=261, 2020. Accessed: 2025-05-19.

[2] AppTek. Apptek mediasphere® - ai-enabled global media intelligence. https://www.apptek.ai/solutions/apptek-mediasphere-r-ai-enabled-global-media-monitoring, 2023. Accessed: 2025-05-22.

[3] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32897–32912. Curran Associates, Inc., 2022.

[4] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022.

[5] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual clip. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France, June 2022. European Language Resources Association.

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

[7] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 215–223, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[9] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond english-centric multilingual machine translation. *J. Mach. Learn. Res.*, 22(1), January 2021.

[10] Li Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 178–178, 2004.

[11] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12976–12985, June 2021.

[12] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, Zongzheng Xi, Yueqian Yang, Anwen Hu, Jinming Zhao, Ruichen Li, Yida Zhao, Liang Zhang, Yuqing Song, Xin Hong, Wanqing Cui, Danyang Hou, Yingyan Li, Junyi Li, Peiyu Liu, Zheng Gong, Chuhao Jin, Yuchong Sun, Shizhe Chen, Zhiwu Lu, Zhicheng Dou, Qin Jin, Yanyan Lan, Wayne Xin Zhao, Ruihua Song, and Ji-Rong Wen. Wenlan: Bridging vision and language by large-scale multi-modal pre-training, 2021.

[13] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: Multimodal, multitask retrieval across languages, 2021.

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR, 18–24 Jul 2021.

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

[16] Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Dong Hyeon Jeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, Na-Hyeon Ryu, Kang Min Yoo, Minsuk Chang, Soobin Suh, Sookyo In, Jinseong Park, Kyungduk Kim, Hiun Kim, Jisu Jeong, Yong Goo Yeo, Donghoon Ham, Dongju Park, Min Young Lee, Jaewook Kang, Inho Kang, Jung-Woo Ha, Woomyoung Park, and Nako Sung. What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers, 2021.

[17] Ildoo Kim, Gunsoo Han, Jiyeon Ham, and Woonhyuk Baek. Kogpt: Kakaobrain korean(hangul) generative pre-trained transformer. `https://github.com/kakaobrain/kogpt`, 2021.

[18] Jake Tae Kim. Koclip: Contrastive vision-language pretraining in korean. `https://github.com/jaketae/koclip`, 2022. Accessed: 2025-05-20.

[19] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[20] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc., 2021.

[21] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: Towards unified-modal understanding and generation via cross-modal contrastive learning. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2592–2607, Online, August 2021. Association for Computational Linguistics.

[22] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022.

[23] Robert Litschko, Goran Glavaš, Simone Paolo Ponzetto, and Ivan Vulić. Unsupervised cross-lingual information retrieval using monolingual data only. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1253–1256, New York, NY, USA, 2018. Association for Computing Machinery.

[24] Han Liu, Yangyang Guo, Jianhua Yin, Zan Gao, and Liqiang Nie. Review polarity-wise recommender, 2022.

[25] Peng Liu, Lemei Zhang, and Jon Atle Gulla. Multilingual review-aware deep recommender system via aspect-based sentiment analysis. 39(2), January 2021.

[26] Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jung-Woo Ha, and Kyunghyun Cho. Klue: Korean language understanding evaluation, 2021.

[27] 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. Learning transferable visual models from natural language supervision, 2021.

[28] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15638–15650, June 2022.

[29] Stefan Smeu, Elisabeta Oneata, and Dan Oneata. Declip: Decoding clip representations for deepfake localization, 2024.

[30] QuoQA-NLP Team. Kocc3m: Korean conceptual captions 3m dataset. `https://huggingface.co/datasets/QuoQA-NLP/KoCC3M`, 2024. Accessed: 2025-05-20.

[31] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Multilingual e5 text embeddings: A technical report, 2024.

[32] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[33] Zehan Wang, Yang Zhao, Xize Cheng, Haifeng Huang, Jiageng Liu, Li Tang, Linjun Li, Yongqi Wang, Aoxiong Yin, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations, 2023.

[34] Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou. Xgpt: Cross-modal generative pre-training for image captioning. In Lu Wang, Yansong Feng, Yu Hong, and Ruifang He, editors, *Natural Language Processing and Chinese Computing*, pages 786–797, Cham, 2021. Springer International Publishing.

[35] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4514–4528. Curran Associates, Inc., 2021.

[36] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese, 2023.

A Team contributions

All team members are equally contributed.

- *Seunggi, Moon* : Image dataset curation, Architecture figure, Evaluation
- *Yujin, Sung* : Korean text dataset curation, Train (image, text memory), Evaluation
- *Donghyun, Shin* : English text dataset curation, Train (Overall), Evaluation
- *Dahyun, Chung* : English text dataset curation, Paper writing, Analysis

B Appendix

B.1 Ablation Studies

Multilingual Encoder	Params	I→K			K→I		
		R@1	R@5	R@10	R@1	R@5	R@10
E5-base[31]	110M	0.3228	0.7890	0.8828	0.3812	0.7812	0.9
MiniLM-L12[32]	33.4M	0.6563	0.9297	0.9687	0.5859	0.8859	0.9453

Table 3: **Ablation study on multilingual text encoders.** "Params" denotes the number of model parameters. I→K denotes Image-to-Korean retrieval and K→I denotes Korean-to-Image retrieval. We compare E5[31] and MiniLM[32] under identical training settings. Best scores are highlighted.

To assess the impact of multilingual encoder selection, we conduct an ablation study comparing two widely used language models: E5-base[31] and MiniLM-L12[32]. These encoders differ in architecture size, training objective, and multilingual representation quality. To isolate the effect of the encoder, we replace only the multilingual encoder while keeping the CLIP encoders and all other training configurations fixed. The projection heads $f_1(\cdot)$, $f_2(\cdot)$ are retrained for each variant. In Table 3, we evaluate the impact of each encoder on two Korean-visual downstream tasks: Image-to-Korean and Korean-to-Image retrieval on MSCOCO-Kor[1], using Recall@K as the metric. Despite its smaller size, MiniLM-L12 consistently outperforms E5-base across all retrieval metrics. The strong performance of MiniLM-L12 supports our choice of encoder in the main experiments and highlights its suitability for cross-lingual vision-language tasks.

B.2 Inference time

Method	Translation Time (ms)	Encoding Time (ms)	Total Inference Time (ms)
CLIP + Translator [9]	350.2	27.4	**377.6**
HanCLIP (Ours)	×	36.3	**36.3**

Table 4: **Inference time comparison.** We report average per-query inference time in milliseconds, including translation and encoding.

Following prior work such as mCLIP [6], which measured inference latency for translation-based versus multilingual approaches, a common baseline for Korean–image retrieval is to first translate Korean text into English, then apply a pre-trained CLIP model. While simple to implement, this two-stage pipeline introduces significant inference latency, particularly due to the costly translation step, which hinders real-time or resource-constrained applications. We compare this pipeline with *HanCLIP*, which directly encodes Korean queries using a multilingual encoder. Inference time was measured on a single NVIDIA TitanXp GPU. For the baseline, we include both translation and CLIP encoding times; for *HanCLIP*, we measure multilingual encoding and projection. As shown in Table 4, *HanCLIP* achieves more than a 10 times improvement in inference speed (36.3 ms vs. 377.6 ms), dramatically enhancing efficiency by eliminating the translation overhead entirely.

Module	Block	C_{in}	C_{out}
Projector	Linear	384	768
	BatchNorm1D	768	768
	ReLU	-	-
	Linear	768	512

Table 5: **Architecture of projection layer**

B.3 Architecture of projection layer

The two projectors, $f_1(\cdot)$ and $f_2(\cdot)$, share the same architecture and are implemented as multi-layer perceptrons (MLPs) composed of two linear layers. As described in Table 5, the first linear layer expands the input dimensionality from 384 to 768, followed by a BatchNorm1D layer and a ReLU activation function. The second linear layer projects the hidden representation down to 768 dimensions. Notably, f_1 and f_2 differ only in the configuration of the BatchNorm1D layer: while f_1 normalizes features with dimension 768, f_2 retains the same dimension without further expansion. This consistent projection structure allows both Korean and image modalities to be mapped into a shared 512-dimensional embedding space.

B.4 Qualitative Korean-image retrieval results

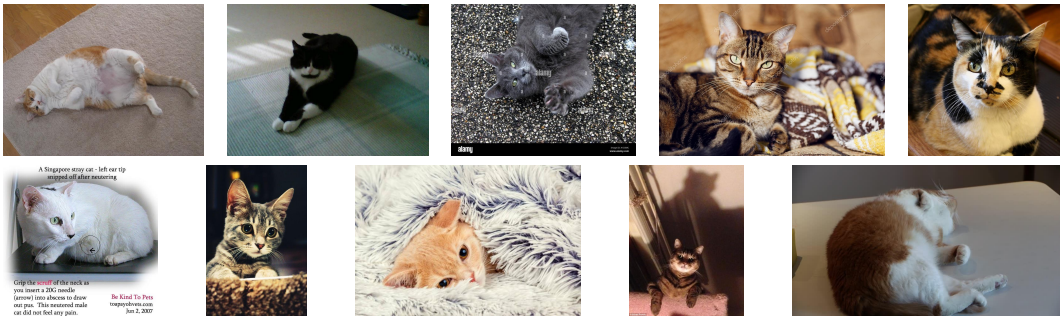

Figure 4: **Retrieval results for the Korean query "누워있는 고양이".**