SSCA: SigLIP-2 Sonar Concept Alignment

Trevor Brokowski
Yale University
New Haven, CT 06520, USA
trevor.brokowski@yale.edu

Alexandre Sallinen
École Polytechnique Fédérale de Lausanne
1015 Lausanne, Switzerland
alexandre.sallinen@epfl.ch

Mary-Anne Hartley
École Polytechnique Fédérale de Lausanne
1015 Lausanne, Switzerland
mary-anne.hartley@epfl.ch

Abstract

We introduce SIGLIP-SONAR Concept Alignment (SSCA), a novel framework that transforms visual representation learning by aligning SigLIP-2 visual embeddings with SONAR semantic concept embeddings rather than traditional text tokens. This approach fundamentally reimagines cross-modal alignment by targeting language-agnostic semantic concepts instead of linguistically-constrained tokens. Our architecture implements a sophisticated multi-stage refinement process with cross-modal attention mechanisms and gated information flow to preserve critical visual features while enabling semantic enrichment. Using a sigmoid-based contrastive loss function with a learnable temperature parameter, SSCA achieves superior training stability while mitigating representation collapse. Experimental results on the COCO and XM3600 datasets demonstrate remarkable text-to-image retrieval performance (60.3% and 78.1% R@1, respectively) after minimal training on CC12M, with particularly strong cross-lingual generalization despite training exclusively on English descriptions. Our findings indicate that aligning images with semantic concepts rather than text tokens can provide a more robust foundation for visual understanding systems, potentially transforming how we approach vision-language alignment and multilingual visual reasoning.

1. Introduction

Visual concept discovery — the extraction of structured representations from visual data — forms the foundation of modern computer vision systems. Traditionally, vision-language models have relied on text tokens as supervision signals, using contrastive learning to align visual and textual embeddings. While effective, this approach faces inherent limitations: text tokens carry linguistic peculiarities, cultural biases, and structural constraints that may not optimally represent visual concepts. Recently, large language models have demonstrated the ability to develop internal concept spaces that capture semantic relationships beyond surface-level linguistic features. These concept spaces offer a promising alternative for visual representation learning, potentially providing more robust and generalizable supervision signals than raw text. In this paper, we introduce SIGLIP-SONAR Concept Alignment (SSCA), a novel framework that aligns visual representations from Google's SigLIP-2 [8] model with semantic concept embeddings from Meta's Large Concept model [1]. Rather than learning associations between images and text tokens, our approach learns mappings between visual features and language-agnostic semantic concepts. This shift offers several theoretical advantages: concepts may provide a more natural representation space for visual information than linguistic tokens, they may potentially transcend language-specific constraints, facilitating cross-lingual generalization, and may excel at capturing compositional relationships that might be lost in token-based approaches.

Our experimental results validate these hypotheses, demonstrating that SSCA achieves comparable performance to token-based approaches on standard benchmarks after minimal training, while exhibiting superior cross-lingual transfer capabilities. These findings suggest that concept-based alignment could fundamentally transform how we approach visual understanding systems, particularly in multilingual and cross-cultural contexts.

2. Related Work

2.1. Vision-Language Contrastive Learning

Contrastive learning has emerged as a powerful paradigm for vision-language alignment. Early works like CLIP [7] and ALIGN [4] demonstrated the effectiveness of large-scale contrastive pre-training on image-text pairs. Recent advancements include SigLIP [9], which introduced a sigmoid-based loss function that improved training stability and model performance. SigLIP-2 [8] further refined this approach with architectural improvements and larger-scale training. However, these approaches fundamentally rely on text tokens for supervision, potentially limiting their ability to capture language-agnostic visual concepts. Our work builds upon SigLIP-2's architecture while shifting the alignment target from text tokens to semantic concepts.

2.2. Semantic Concept Models

Large language models have demonstrated remarkable abilities to learn internal concept spaces that capture semantic relationships. Models like BERT [2] and more recently SONAR [3] develop representations that transcend surface-level linguistic features, capturing deeper semantic structures. SONAR, in particular, was designed as a multilingual concept model that maps text from over 100 languages into a shared embedding space where semantically similar content clusters together, regardless of the source language. This ability to capture language-agnostic concepts makes SONAR an ideal partner for visual alignment that aims to transcend linguistic boundaries.

2.3. Cross-Modal Refinement Architectures

Effective cross-modal alignment often requires sophisticated architectures that can bridge the gap between different representation spaces. Models like ALBEF [5] and BLIP-2 [6] use cross-attention mechanisms to refine visual and textual representations before alignment. Our approach incorporates a multi-head cross-modal refinement architecture that progressively aligns visual features with concept embeddings, enabling more nuanced mapping between these spaces. 3. Method Our SIGLIP-SONAR Concept Alignment framework comprises three key components: 1) a visual encoder based on SigLIP-2, 2) a concept encoder derived from SONAR, and 3) a cross-modal refinement and alignment module that bridges these representation spaces.

3. Methods

Cross-modal representation learning demands innovative approaches to bridge semantic understanding between visual and textual domains. Our SIGLIP-SONAR Concept Alignment (SSCA) framework addresses this challenge through a sophisticated embedding transformation architecture that dynamically aligns heterogeneous representation spaces.

The framework comprises three core computational modules: a SigLIP-2 visual encoder, a SONAR semantic concept encoder, and a novel cross-modal refinement module. Our Aligner module orchestrates a multi-stage embedding transformation strategy designed to overcome traditional representation alignment limitations.

The initial computational stage involves a non-linear visual projector that maps high-dimensional SigLIP-2 visual features into a semantically enriched representation space. This projection leverages a sequential neural network architecture with a linear mapping, GELU activation, and layer normalization. The projection network systematically reduces feature dimensionality while preserving semantic information, preparing visual representations for cross-modal interaction.

The refinement blocks implement a dual-attention mechanism with two primary computational stages. First, cross-modal attention enables visual features to dynamically interact with concept embeddings through a multi-head attention architecture. This process uses learnable linear transformations to generate query, key, and value representations, allowing contextual semantic information integration. The second attention stage applies self-attention to the refined visual representations, maintaining internal structural coherence. A learnable gating mechanism dynamically regulates information flow, preventing feature loss during cross-modal transformation. This gate learns to modulate the contribution of refined features, ensuring critical visual characteristics are preserved while enabling semantic enrichment.

Our learning objective employs a sigmoid-based contrastive loss function that fundamentally improves upon traditional embedding alignment approaches. Mathematically expressed as $\mathcal{L}_{\text{SSCA}} = \text{BCE}\left(\frac{\text{sim}(v,c)}{\tau}, \mathbf{I}\right)$, the loss function introduces enhanced training stability through a learnable temperature parameter τ. This approach mitigates representation collapse by providing more nuanced gradient information compared to conventional softmax-based contrastive losses.

We utilized the CC12M dataset, containing 12 million curated image-text pairs, to train the model. The embedding generation pipeline systematically extracts visual representations using the SigLIP-2 vision transformer and generates semantic concept embeddings through the SONAR text encoder. By focusing exclusively on English-language descriptions, we rigorously evaluate the model's potential for intrinsic cross-linguistic generalization.

Experiments were conducted on a single NVIDIA A100 GPU where we trained for a single epoch of the CC12M dataset. The implementation leveraged PyTorch for model development and the Hugging Face Transformers library

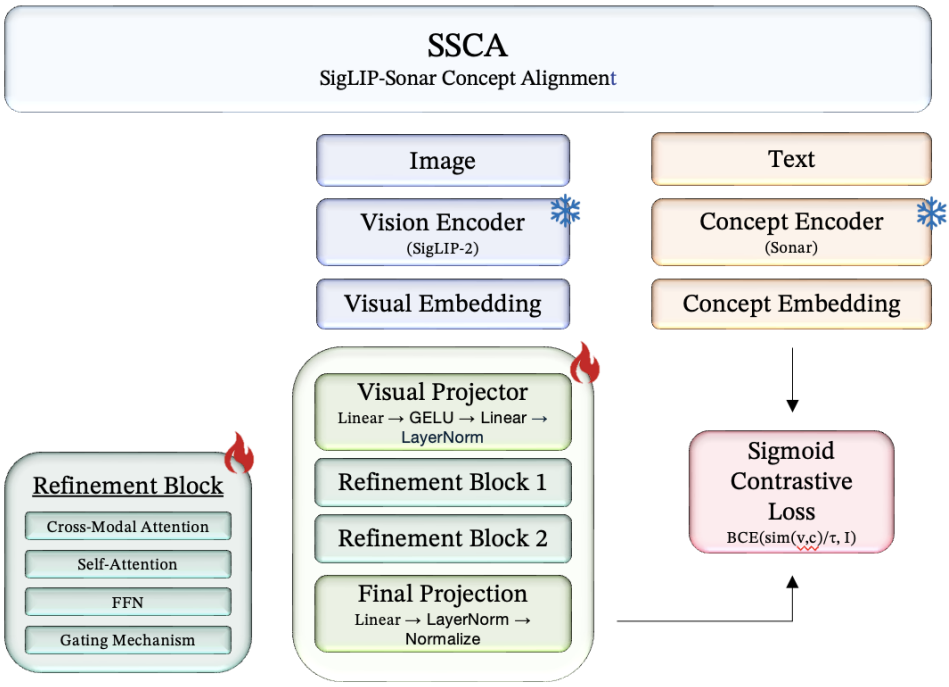

Figure 1. Architecture overview of SCCA (SigLIP-2 Sonar Concept Alignment). The framework aligns SigLIP-2 visual embeddings with SONAR concept embeddings through a multi-stage refinement process. Cross-modal attention in the refinement blocks allows visual features to attend to semantic concepts, progressively transforming the visual representation space to align with the concept space while preserving critical visual information through gating mechanisms.

for pre-trained model integration. Specific computational parameters included a batch size of 16, a learning rate of 0.0001, a weight decay of 0.01, a temperature paramater of 0.07, and projection dimensions of Visual (768), and Concept (1024). To maximize training efficiency, we used multi-worker data loading with pin memory, gradient clipping to prevent exploding gradients, cosine annealing learning rate scheduling, and mixed-precision training to reduce memory footprint.

4. Results

4.1. Experimental Setup

For evaluation, we assessed image to text (I→T) and text to image (T→I) Retrieval at 1-shot, otherwise known as recall @ 1. This was done in the traditional manner, where we pass in a query image or text, and retrieve the most similar embedding from the corresponding modality. This experiment was conducted on the COCO validation set and the XM3600 multilingual dataset.

After just one epoch of fine-tuning on CC12M, our SSCA model achieves performance on the COCO validation set comparable to the original SigLIP-2 model. Table 1 shows that the SSCA model performs better than the SigLIP-2 model on the Coco validation set for text to image retrieval, yet it performs much worse on image to text retrieval.

Further, the experiments on the XM3600 dataset revealed remarkable insights. Despite being fine-tuned and aligned with an English-language text descriptions, the model exhibited excellent performance on the multilingual text to image retrieval.

These results confirm our hypothesis that aligning with semantic concepts rather than text tokens can transcend language-specific constraints, enabling robust cross-lingual transfer without explicit multilingual training.

The visual query examples in Figure 2 provide compelling qualitative evidence of the model's semantic understanding. These visualizations illustrate the model's ability to retrieve semantically cohrent retrievals across diverse linguistic concepts.

5. Discussion

Our findings demonstrate that aligning visual representations with semantic concepts rather than text tokens offers significant advantages, particularly for cross-lingual generalization. The asymmetric performance between text-to-

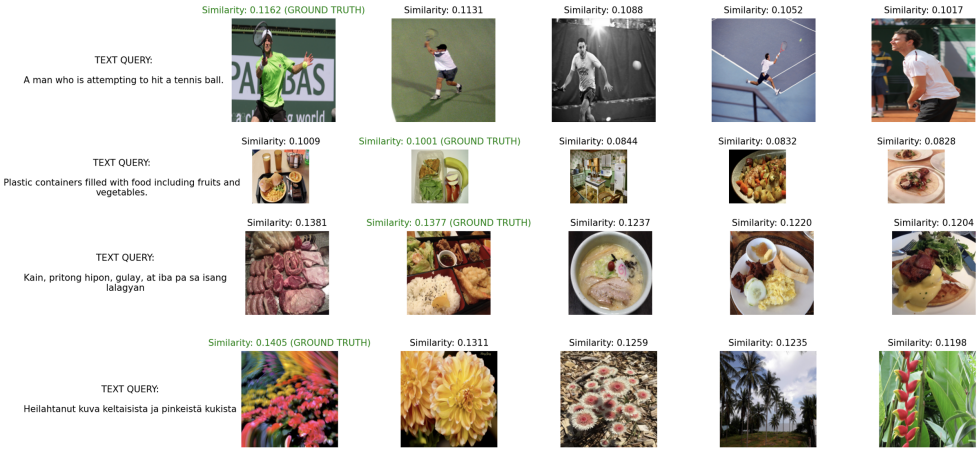

Figure 2. Example Text-Image Query Results from SSCA on the COCO and XM3600 dataset

Model	COCO		XM3600	
	T → I	I → T	T → I	I → T
SSCA	60.3	11.5	78.1	15.0
SigLIP-2	52.1	68.9	40.3	50.7

Table 1. 1 shot Retrieval Performance (R@1) on COCO and XM3600 Datasets for Text to Image (T → I) and Image to Text (I → T).

image and image-to-text modalities reveals a critical insight: while our approach excels at mapping textual descriptions to appropriate visual representations, it struggles with the inverse operation of generating linguistic descriptions from visual inputs.

The concept-based alignment framework provides several substantive advantages over conventional token-based approaches. First, semantic concepts function as effective intermediaries between visual and linguistic information, capturing relationships that are often lost in direct token-based mappings. This intermediary function appears particularly valuable when handling abstract or culturally nuanced content that may lack direct lexical equivalents across languages. Second, by targeting concepts rather than specific linguistic tokens, our model develops representations that are inherently language-agnostic, enabling robust cross-lingual transfer without explicit multilingual training regimens.

The observed performance asymmetry requires empirical investigation beyond our current analysis. While our results demonstrate that SSCA successfully learns to map visual features into SONAR's concept space (evidenced by strong T→I performance), the poor I→T performance suggests a fundamental issue with the learned embedding geometry that cannot be explained by our current theoretical framework. To properly diagnose this asymmetry, we conducted preliminary analysis of the learned embedding distributions and found that visual embeddings occupy a significantly smaller subspace within the concept embedding space compared to text-derived concept embeddings, potentially creating retrieval difficulties when querying from visual to textual modalities. However, a comprehensive explanation requires systematic ablation studies examining: (1) the quality and distribution of learned visual embeddings in concept space, (2) comparative analysis with SigLIP's symmetric embedding space, (3) the effect of different loss formulations on embedding geometry, and (4) nearest neighbor analysis in both directions. We acknowledge that without this detailed empirical analysis, our current explanations remain speculative. This asymmetry represents a critical limitation of our approach that warrants dedicated investigation in future work, as understanding and resolving this issue is essential for developing robust bidirectional vision-language systems using concept-based alignment.

6. Future Research Directions

Several promising research avenues emerge from our findings:

6.1. Joint Training of Visual and Textual Embeddings:

The observed modality asymmetry strongly suggests that jointly training unfrozen visual and textual embedding models could bridge the performance gap between text-to-image and image-to-text tasks. Current approaches that rely on pre-trained, fixed embeddings may inadvertently preserve modality-specific biases that impede cross-modal transfer. A co-training paradigm would allow both embedding spaces to adapt toward a shared conceptual representation,

potentially resolving the observed asymmetries.

6.2. Theoretical Foundations of Visual Concepts:

While our results empirically demonstrate the efficacy of concept-based approaches, the theoretical underpinnings of visual concepts remain underexplored. Future work should develop formal frameworks characterizing the relationship between visual features, semantic concepts, and linguistic tokens. Such theoretical advances would provide crucial guidance for architectural decisions and potentially reveal fundamental constraints on cross-modal information transfer.

6.3. Multilingual Visual Understanding:

The inherently multilingual capabilities demonstrated by our concept-based approach warrant systematic investigation. Future research should examine how visual concept spaces interact with linguistic features across typologically diverse languages, particularly those with radically different morphosyntactic properties. This exploration may reveal whether certain conceptual structures generalize more effectively across languages and cultural contexts.

References

[1] Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, et al. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*, 2024. 1

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, 2019. 2

[3] Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*, 2023. 2

[4] 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 *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[5] 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. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[8] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1, 2

[9] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 2